

MÉTODOS DE ANÁLISIS ESTADÍSTICO: DESDE LO DESCRIPTIVO HASTA LO INFERENCIAL


ISBN: 978-9915-698-07-6

Vicky Leonor Alata Linares - María Luz Maldonado Peña - Denisse Montalvan Alburqueque - Michaels Pedro Mejia Lagos - Jose Carlos Fiestas Zevallos - Hector Fidel Bejarano Benites - Luis Alberto Aguirre Bazán

Métodos de análisis estadístico: Desde lo descriptivo hasta lo inferencial

Vicky Leonor Alata Linares, María Luz Maldonado Peña, Denisse Montalvan Alburqueque, Michaels Pedro Mejia Lagos, Jose Carlos Fiestas Zevallos, Héctor Fidel Bejarano Benites, Luis Alberto Aguirre Bazán

© Vicky Leonor Alata Linares, María Luz Maldonado Peña, Denisse Montalvan Alburqueque, Michaels Pedro Mejia Lagos, Jose Carlos Fiestas Zevallos, Héctor Fidel Bejarano Benites, Luis Alberto Aguirre Bazán, 2025

Primera edición: Mayo, 2025

Editado por:

Editorial Mar Caribe

www.editorialmarcaribe.es

Av. General Flores 547, Colonia, Colonia-Uruguay.

Diseño de portada: Yelitza Sánchez Cáceres

Libro electrónico disponible en:
https://editorialmarcaribe.es/ark:/10951/isbn.9789915698076

Formato: electrónico

ISBN: 978-9915-698-07-6

ARK: ark:/10951/isbn.9789915698076

URN: URN:ISBN:978-9915-698-07-6

Atribución/Reconocimiento-NoComercial 4.0 Internacional:

Los autores pueden autorizar al público en general a reutilizar sus obras únicamente con fines no lucrativos, los lectores pueden utilizar una obra para generar otra, siempre que se dé crédito a la investigación, y conceden al editor el derecho a publicar primero su ensayo bajo los términos de la licencia CC BY-NC 4.0.

Editorial Mar Caribe, firmante Nº 795 de 12.08.2024 de la Declaración de Berlín:

"... Nos sentimos obligados a abordar los retos de Internet como medio funcional emergente para la distribución del conocimiento. Obviamente, estos avances pueden modificar significativamente la naturaleza de la publicación científica, así como el actual sistema de garantía de calidad...." (Max Planck Society, ed. 2003., pp. 152-153).

Editorial Mar Caribe-Miembro de OASPA:

Como miembro de la Open Access Scholarly Publishing Association, apoyamos el acceso abierto de acuerdo con el código de conducta, transparencia y mejores prácticas de OASPA para la publicación de libros académicos y de investigación. Estamos comprometidos con los más altos estándares editoriales en ética y deontología, bajo la premisa de «Ciencia Abierta en América Latina y el Caribe».








Editorial Mar Caribe

Métodos de análisis estadístico: Desde lo descriptivo hasta lo inferencial

Colonia, Uruguay

2025

Sobre los autores y la publicación

Vicky Leonor Alata Linares

valatal@usmp.pe

https://orcid.org/0000-0003-1897-5757

Universidad de San Martín de Porres, Perú

María Luz Maldonado Peña

mmaldonadop@usmp.pe

https://orcid.org/0000-0002-3143-769X

Universidad de San Martín de Porres, Perú

Denisse Montalvan Alburqueque

dmontalvan@cientifica.edu.pe

https://orcid.org/0009-0007-6828-1178

Universidad Científica del Sur, Perú

Michaels Pedro Mejia Lagos

mmejia@cientifica.edu.pe

https://orcid.org/0009-0009-9863-7184

Universidad Científica del Sur, Perú

Héctor Fidel Bejarano Benites

hbejarano@cientifica.edu.pe

https://orcid.org/0000-0003-2047-4425

Universidad Científica del Sur, Perú

Jose Carlos Fiestas Zevallos

jfiestasz@unp.edu.pe

https://orcid.org/0009-0008-7860-5911

Universidad Nacional de Piura, Perú

Luis Alberto Aguirre Bazán

laguirreb@unitru.edu.pe

https://orcid.org/0000-0002-5642-1213

Universidad Nacional de Trujillo, Perú

Libro resultado de investigación:

Publicación original e inédita, cuyo contenido es el resultado de un proceso de investigación llevado a cabo con anterioridad a su publicación, ha sido sometida a una revisión externa por pares a doble ciego, el libro ha sido seleccionado por su calidad científica y porque contribuye significativamente al área de conocimiento e ilustra una investigación completamente desarrollada y finalizada. Además, la publicación ha pasado por un proceso editorial que garantiza su normalización bibliográfica y usabilidad.

Sugerencia de citación:

Alata, V.L., Maldonado, M.L., Montalvan, D., Mejia, M.P., Bejarano, H.F., Fiestas, J.C., y Aguirre, L.A. (2025). *Métodos de análisis estadístico: Desde lo descriptivo hasta lo inferencial*. Colonia del Sacramento: Editorial Mar Caribe. https://editorialmarcaribe.es/ark:/10951/isbn.9789915698076

Índice

Introducción

Los métodos descriptivos son fundamentales en el análisis estadístico, ya que permiten resumir, organizar y presentar de manera clara y concisa los datos recopilados. A través de estas técnicas, es posible obtener una visión general de las características de un conjunto de datos sin realizar inferencias sobre una población más amplia. En este libro, exploraremos la definición y el propósito de los métodos descriptivos, así como las medidas de tendencia central y de dispersión que se utilizan comúnmente.

Los métodos descriptivos tienen como objetivo principal proporcionar un resumen de las características esenciales de un conjunto de datos, para presentar información de manera que sea fácilmente comprensible, facilitando la identificación de patrones, tendencias y anomalías. Al utilizar técnicas descriptivas, se pueden obtener apreciaciones valiosas que guían decisiones informadas y ayudan en la comunicación de resultados a audiencias no especializadas.

Por otra parte, los métodos inferenciales son una parte fundamental del análisis estadístico, ya que permiten a los investigadores realizar generalizaciones sobre una población a partir de una muestra. A diferencia de los métodos descriptivos, que se centran en resumir y describir los datos observados, los métodos inferenciales se utilizan para hacer predicciones y tomar decisiones basadas en esos datos. Además, exploramos el concepto de inferencia estadística, las pruebas de hipótesis y la importancia de los intervalos de confianza.

La inferencia estadística es el proceso mediante el cual se extraen conclusiones sobre una población a partir de los resultados obtenidos de una muestra. Este enfoque se basa en la premisa de que una muestra representativa puede ofrecer información valiosa sobre el comportamiento y las características de la población de interés. La inferencia estadística permite a los investigadores no solo describir los datos, sino también hacer afirmaciones más amplias y fundamentadas, lo que es esencial en campos como la medicina, la sociología y la economía.

A lo largo del texto escrito, se ahonda en métodos no paramétricos, que son técnicas estadísticas que no requieren suposiciones sobre la distribución de

los datos. A diferencia de los métodos paramétricos, que asumen que los datos se distribuyen de una manera específica (por ejemplo, normalmente), los métodos no paramétricos son más flexibles y pueden aplicarse a una variedad de situaciones en las que los supuestos de normalidad no se cumplen. Esto los convierte en herramientas valiosas en el análisis de datos, especialmente cuando se trabaja con muestras pequeñas o cuando los datos presentan características no convencionales, como escalas ordinales o distribuciones sesgadas.

Otra característica importante de los métodos no paramétricos es que suelen ser menos sensibles a valores atípicos. Esto significa que su rendimiento no se ve tan afectado por datos extremos como sucede con los métodos paramétricos, lo que los hace más robustos en la práctica. Además, a menudo son más simples de interpretar, ya que se centran en rangos y frecuencias en lugar de en parámetros estadísticos complejos.

En este sentido, las medidas de dispersión son vitales para evaluar la variabilidad y la confiabilidad de los datos, características que son esenciales en el análisis de resultados. Por otro lado, los métodos inferenciales permiten a los investigadores hacer afirmaciones y predicciones más allá de los datos observados. La formulación de hipótesis y la construcción de intervalos de confianza son prácticas que fortalecen la validez de los resultados, proporcionando un marco para evaluar la probabilidad de que ciertos hallazgos sean el resultado de la casualidad. Estas herramientas son indispensables para avanzar en la investigación científica y en la práctica profesional, ya que permiten la validación de teorías y la implementación de políticas basadas en evidencia.

Los autores hacen énfasis en métodos estadísticos, interpretación e inferencia de datos, tal que, asumen que estos datos se distribuyen de una manera específica, más flexibles y que pueden aplicarse a una variedad de situaciones en las que los supuestos de normalidad no se cumplen. Por ende, el objetivo de investigación es, conceptualizar métodos estadísticos descriptivos e inferenciales, a través de pruebas de hipótesis, intervalos de confianza, probabilidades, entre otros, no solo para resumir y describir los datos observados, sino para hacer predicciones y tomar decisiones basadas en esos datos.

Se invita a los lectores a profundizar en métodos paramétricos o no, desde la flexibilidad y robustez, lo que lo convierte en opciones ideales para el análisis de datos que no siguen distribuciones normales o que presentan escalas

ordinales. Esto resalta la importancia del texto escrito, de contar con una variedad de enfoques estadísticos que se adapten a las necesidades específicas de cada estudio.

Capítulo I

Funciones de probabilidad, curtosis, asimetría y correlación biserial

1.1 Entendiendo las Funciones de Probabilidad: Densidad, Distribución Acumulativa e Inversa

Las funciones de probabilidad son herramientas fundamentales en la estadística y la teoría de la probabilidad, ya que nos posibilitan modelar y entender el comportamiento de fenómenos aleatorios. En el ámbito de la estadística, es común encontrarse con tres conceptos clave que forman la base de la probabilidad: la función de densidad de probabilidad (PDF), la función de distribución acumulativa (CDF) y la función de distribución acumulativa inversa. Cada una de estas funciones tiene características y aplicaciones específicas que ayudan a los investigadores y analistas a interpretar datos y realizar inferencias sobre poblaciones a partir de muestras.

La función de densidad de probabilidad (PDF) se utiliza para describir la probabilidad relativa de que una variable aleatoria continua tome un valor específico. A diferencia de las variables aleatorias discretas, donde la probabilidad de ocurrencia de un resultado específico consigue ser directamente calculada, en el caso de las variables continuas, la PDF proporciona una forma de calcular la probabilidad de que la variable caiga dentro de un intervalo determinado (Martinson, 2018). La integral de la PDF en todo su dominio es igual a 1, garantizando que la probabilidad total esté debidamente normalizada.

La función de distribución acumulativa (CDF) es otra herramienta crucial que complementa a la PDF. Definida como la probabilidad de que una variable aleatoria tome un valor menor o igual a un número determinado, la CDF proporciona una forma de evaluar la probabilidad acumulativa de eventos. Este concepto posibilita a los analistas visualizar cómo se distribuyen las probabilidades a lo largo del rango de valores posibles y es especialmente útil al trabajar con datos en diferentes contextos, como en la evaluación de riesgos o en estudios de calidad.

La función de distribución acumulativa inversa es el concepto que nos posibilita obtener el valor de la variable aleatoria correspondiente a una probabilidad acumulada dada. Esta función es fundamental en situaciones donde se requiere desinvertir la CDF para encontrar percentiles o cuantiles. En la generación de números aleatorios que siguen una distribución específica, la función inversa se utiliza para transformar una variable aleatoria uniforme en una variable aleatoria que sigue la distribución deseada. En síntesis, la comprensión de estas funciones es esencial para el análisis y la interpretación de datos en múltiples disciplinas, desde la ingeniería hasta las ciencias sociales.

La función de densidad de probabilidad (PDF, por sus siglas en inglés) es un concepto fundamental en la teoría de la probabilidad y la estadística. Se utiliza para describir el comportamiento de variables aleatorias continuas, proporcionando una forma de caracterizar la probabilidad de que una variable tome un valor específico dentro de un intervalo determinado. Una PDF tiene varias características importantes que la definen:

i. *No Negatividad:* La función de densidad de probabilidad nunca consigue ser negativa. Esto significa que para cualquier valor x, se tiene que cumplir que $f(x) \geq 0$.

ii. *Integral Igual a Uno:* La integral de la PDF sobre todo su dominio debe ser igual a uno, lo que refleja la certeza de que la variable aleatoria tomará algún valor en el espacio muestral. Esto se expresa matemáticamente como:

$$\int_{-\infty}^{\infty} f(x) \, dx = 1$$

iii. *Probabilidad en Intervalos:* La probabilidad de que una variable aleatoria X caiga dentro de un intervalo $[a, b]$ se calcula como la integral de la PDF en ese intervalo:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx$$

iv. *No Proporcionalidad:* A diferencia de las funciones de probabilidad discretas, donde la probabilidad se asigna a valores específicos, en la PDF, la probabilidad de que X tome un valor exacto es cero. Esto se debe a que hay infinitos puntos en un intervalo continuo.

Existen varias funciones de densidad de probabilidad comunes que se utilizan en la práctica:

i. *Distribución Normal:* Es una de las PDF más conocidas y se caracteriza por su forma de campana. Tiene dos parámetros: la media μ y la desviación estándar σ. Su función de densidad se define como:

$$
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}
$$

ii. *Distribución Exponencial:* Utilizada para modelar el tiempo entre eventos en un proceso de Poisson. Su PDF se define como:

$$
f(x) = \lambda e^{-\lambda x} \quad \text{para } x \geq 0
$$

donde λ es la tasa de ocurrencia.

iii. *Distribución Uniforme:* En este caso, todos los intervalos del mismo tamaño tienen la misma probabilidad. La PDF de una distribución uniforme continua en el intervalo $[a, b]$ es:

$$
f(x) = \frac{1}{b-a} \quad \text{para } a \leq x \leq b
$$

La función de densidad de probabilidad tiene múltiples aplicaciones en estadística:

i. *Modelado de Datos:* Las PDF son esenciales para modelar fenómenos naturales y sociales, permitiendo a los investigadores describir y predecir comportamientos.

ii. *Inferencia Estadística:* Muchas técnicas de inferencia estadística, como la estimación de parámetros y la prueba de hipótesis, se fundamentan en suposiciones sobre la forma de la PDF.

iii. *Simulación:* En métodos de simulación, como el muestreo de Monte Carlo, se utilizan PDF para generar valores aleatorios que siguen una distribución específica.

iv. *Análisis de Riesgos:* En el análisis de riesgos financieros y actuariales, las PDF posibilitan cuantificar la probabilidad de eventos adversos y evaluar la exposición al riesgo.

En otras palabras, la función de densidad de probabilidad es un pilar fundamental en la estadística y la teoría de la probabilidad, proporcionando una base sólida para el análisis y la interpretación de datos en una variedad de disciplinas.

1.1.1 Función de distribución acumulativa (CDF)

La función de distribución acumulativa (CDF, por sus siglas en inglés) es una herramienta fundamental en la teoría de probabilidad que posibilita describir la distribución de una variable aleatoria. Su definición se basa en la probabilidad de que una variable aleatoria X tome un valor menor o igual a un número específico x (Kuter, 2025). Matemáticamente, se expresa como:

$$F(x) = P(X \leq x)$$

Donde $F(x)$ representa la CDF. Esta función tiene varias propiedades importantes:

i. *Monotonía no decreciente:* La CDF es una función no decreciente, lo que significa que si $a < b$ entonces $F(a) \leq F(b)$.

ii. *Límites:* A medida que x tiende a menos infinito, $F(x)$ tiende a 0, y cuando x tiende a más infinito, $F(x)$ tiende a i. Esto asegura que la CDF abarca todas las probabilidades posibles de la variable aleatoria.

iii. *Continuidad:* Para variables aleatorias continuas, la CDF es continua, mientras que para variables discretas consigue presentar saltos en valores específicos.

La relación entre la función de densidad de probabilidad (PDF) y la función de distribución acumulativa (CDF) es fundamental para entender la distribución de probabilidades. Para una variable aleatoria continua, la CDF se consigue obtener integrando la PDF en el intervalo desde menos infinito hasta x:

$$
F(x) = \int_{-\infty}^{x} f(t) \, dt
$$

Donde $f(t)$ es la PDF de la variable aleatoria X. Por otro lado, la PDF consigue ser derivada de la CDF al calcular la derivada de $F(x)$:

$$
f(x) = \frac{dF(x)}{dx}
$$

Esta relación muestra cómo ambas funciones son interdependientes y proporciona un puente entre la representación acumulativa y la densidad de probabilidad (Kuter, 2025). Para ilustrar el concepto de la función de distribución acumulativa, consideremos algunos ejemplos prácticos:

i. *Distribución Normal:* La CDF de una variable aleatoria normalmente distribuida no consigue expresarse en forma cerrada, pero se consigue calcular utilizando tablas o software estadístico. Esta CDF posibilita determinar la probabilidad de que una observación caiga dentro de un rango específico de valores.

ii. *Distribución Uniforme:* Para una variable aleatoria X que sigue una distribución uniforme en el intervalo $[a, b]$, la CDF se define como:

\[

F(x) =

\begin{cases}

0 & \text{si } x < a \\

\frac{x - a}{b - a} & \text{si } a \leq x < b \\

1 & \text{si } x \geq b

\end{cases}

\]

Esto significa que la probabilidad de que \(X\) sea menor o igual a un valor específico \(x\) es lineal entre \(a\) y \(b\).

iii. *Distribución Exponencial:* Para una variable aleatoria \(X\) con una distribución exponencial con parámetro \(\lambda\), la CDF se expresa como:

\[

F(x) = 1 - e^{-\lambda x} \quad \text{para } x \geq 0

\]

Este ejemplo es común en el modelado de tiempos de espera y eventos aleatorios en procesos de Poisson. Estos ejemplos resaltan la utilidad de la CDF en la evaluación de probabilidades en diferentes contextos y distribuciones, y su importancia en la estadística y la teoría de probabilidad.

1.1.2 Función de distribución acumulativa inversa

La función de distribución acumulativa inversa, a menudo denotada como \(F^{-1}(p) \), es una herramienta fundamental en la teoría de probabilidad y estadística. Su principal función es proporcionar el valor en el dominio de la variable aleatoria correspondiente a un nivel de probabilidad específico. En otras palabras, mientras que la función de distribución acumulativa (CDF) nos dice la probabilidad de que una variable aleatoria tome un valor menor o igual a un cierto número, la función acumulativa inversa nos posibilita encontrar el valor de la variable aleatoria para una probabilidad dada. La función de distribución acumulativa inversa se define como:

\[

F^{-1}(p) = x \quad \text{si} \quad F(x) = p

\]

donde \(p \) es un valor en el intervalo [0, 1] que representa la probabilidad acumulada y \(x \) es el valor correspondiente de la variable aleatoria. Esta función es especialmente útil en la generación de números aleatorios y en el muestreo, ya que posibilita transformar un número aleatorio uniformemente

distribuido en un número que sigue la distribución deseada. Si se desea obtener un número que siga una distribución normal a partir de un número aleatorio que sigue una distribución uniforme, se consigue aplicar la función de distribución acumulativa inversa de la distribución normal para obtener el resultado deseado.

Existen varios métodos para calcular la función de distribución acumulativa inversa, dependiendo de la complejidad de la distribución. Algunos de los métodos más comunes son:

i. *Método analítico:* En algunos casos, la función inversa se consigue derivar algebraicamente. Para distribuciones bien conocidas como la normal o la exponencial, existen fórmulas explícitas que posibilitan calcular \(F^{-1}(p) \) directamente.

ii. *Método numérico:* Cuando no se consigue obtener una solución analítica, se consiguen utilizar métodos numéricos como el método de bisección o el método de Newton-Raphson para aproximar el valor de \(x \) para un \(p \) dado. Estos métodos implican iterar sobre posibles valores de \(x \) hasta que se alcance un nivel de precisión deseado.

iii. *Tablas y software estadístico:* Para muchas distribuciones comunes, se han desarrollado tablas y software que proporcionan directamente los valores de la función acumulativa inversa. Herramientas como R, Python (con bibliotecas como SciPy) y Excel presentan funciones integradas que facilitan este cálculo.

La función de distribución acumulativa inversa tiene numerosas aplicaciones en diversos campos:

i. *Simulación y muestreo:* En el ámbito de la simulación, se utiliza para generar números aleatorios que siguen una distribución específica. Este proceso es esencial en técnicas como el muestreo de Monte Carlo, donde se requiere generar muestras aleatorias de manera eficiente.

ii. *Análisis de riesgos:* En finanzas y análisis de riesgos, la función inversa se utiliza para determinar percentiles de retorno de inversiones, lo que posibilita a los analistas evaluar el riesgo asociado a diferentes niveles de rendimiento.

iii. *Inferencia estadística:* En estudios de inferencia, se consigue utilizar la función inversa para encontrar límites de confianza y realizar pruebas de hipótesis, proporcionando un marco para tomar decisiones basadas en datos.

Como se ha dicho, la función de distribución acumulativa inversa es una herramienta valiosa que, al igual que sus contrapartes, la función de densidad de probabilidad y la función de distribución acumulativa, desempeña un papel crucial en la comprensión y aplicación de conceptos probabilísticos en diversas disciplinas. La densidad de probabilidad (PDF) nos posibilita describir cómo se distribuyen las probabilidades en un conjunto de resultados posibles, ofreciendo una representación clara de los comportamientos de distintas variables aleatorias (Batanero y Díaz 2011). Su importancia se extiende a diversas aplicaciones en estadística y otras disciplinas, donde es fundamental entender la naturaleza de los datos.

Por otro lado, la CDF proporciona una herramienta valiosa para evaluar la probabilidad acumulativa de que una variable aleatoria tome un valor menor o igual a un umbral específico. Esta función nos brinda información esencial sobre la tendencia y la distribución de los datos, sirviendo como base para realizar inferencias y tomar decisiones informadas.

La función de distribución acumulativa inversa es igualmente significativa, ya que posibilita deshacer el proceso de acumulación y acceder a valores específicos de probabilidad. Esto resulta especialmente útil en la generación de números aleatorios y en la realización de simulaciones, donde se requiere conocer el valor correspondiente a una probabilidad dada.

La comprensión y el uso adecuado de la PDF, la CDF y su función inversa son esenciales para el análisis estadístico y la modelización de fenómenos aleatorios. Estudiar y aplicar estas funciones no solo enriquece nuestro conocimiento teórico, sino que al igual mejora nuestra capacidad para aplicar técnicas estadísticas en situaciones prácticas, contribuyendo así al avance de la ciencia y la investigación en diversas áreas.

1.2 Entendiendo la Asimetría y Curtosis: Cálculo e Interpretación en Estadística

En el análisis estadístico, comprender la distribución de los datos es fundamental para la interpretación y la toma de decisiones informadas. Dos conceptos clave que emergen en este contexto son la asimetría y la curtosis. Ambos indicadores proporcionan información valiosa sobre la forma de una distribución y ayudan a identificar patrones, tendencias y anomalías.

La asimetría se refiere a la simetría de la distribución de los datos en relación con su media. Una distribución consigue ser simétrica, lo que significa que los datos se distribuyen de manera uniforme alrededor de la media, o consigue presentar asimetría, donde existen desviaciones en uno de los extremos. Esta característica es crucial, ya que consigue afectar la validez de diversas pruebas estadísticas y la interpretación de los resultados.

Por otro lado, la curtosis se refiere al "apuntamiento" de la distribución. Mientras que la asimetría se ocupa de la inclinación de la distribución, la curtosis se focaliza en la concentración de los datos en torno a la media. Una alta curtosis indica que los datos tienen colas pesadas y un pico pronunciado, mientras que una baja curtosis sugiere una distribución más plana y con colas ligeras. La curtosis proporciona una visión adicional sobre la dispersión de los datos y la probabilidad de observar valores extremos.

La asimetría es un concepto fundamental en la estadística que se refiere a la falta de simetría en la distribución de un conjunto de datos. Cuando se analiza un conjunto de datos, es crucial entender cómo se distribuyen los valores en relación con la media. Una distribución se considera simétrica si los valores se distribuyen de manera uniforme a ambos lados de la media, mientras que una distribución asimétrica muestra una inclinación hacia un lado, lo que consigue influir en la interpretación y el análisis de los datos (Fau y Nabzo, 2020).

La asimetría se define como la medida de la desviación de una distribución de datos respecto a su media. Específicamente, se refiere a la dirección y el grado de la inclinación de la distribución. Por lo general, la asimetría se expresa en términos de un valor numérico que consigue ser positivo, negativo o cero. Un valor de asimetría cero indica una distribución perfectamente simétrica, mientras que un valor positivo indica una asimetría hacia la derecha (más valores extremos en la cola derecha) y un valor negativo indica una asimetría hacia la izquierda (más valores extremos en la cola izquierda). Existen dos tipos principales de asimetría:

i. *Asimetría positiva:* En una distribución con asimetría positiva, la cola de la distribución se extiende más hacia la derecha. Esto significa que hay una mayor cantidad de datos concentrados en el lado izquierdo de la media, mientras que los valores más altos se encuentran en la cola derecha. La distribución de ingresos en una población consigue tener una asimetría positiva, donde la mayoría de las

personas ganan un ingreso bajo, pero un pequeño número tiene ingresos muy altos.

ii. *Asimetría negativa:* Por otro lado, en una distribución con asimetría negativa, la cola de la distribución se extiende hacia la izquierda. Aquí, la mayoría de los datos se concentran en el lado derecho de la media, mientras que los valores más bajos se encuentran en la cola izquierda. Un ejemplo de asimetría negativa podría ser la distribución de edades de jubilación, donde la mayoría de las personas se retiran a una edad avanzada, pero algunas consiguen hacerlo mucho antes.

La asimetría juega un papel crucial en la estadística, ya que consigue afectar la interpretación de los resultados y la selección de métodos estadísticos apropiados para el análisis de datos. Una distribución asimétrica consigue indicar que los datos no se ajustan a las suposiciones de normalidad requeridas por muchos métodos estadísticos, como las pruebas t o el análisis de varianza. Entonces, comprender la asimetría ayuda a los analistas a identificar patrones, tendencias y posibles sesgos en los datos, lo que a su vez consigue influir en la toma de decisiones y en la formulación de hipótesis. Dicho de otro modo, la asimetría es una medida esencial para describir la forma de una distribución de datos y para evaluar su impacto en el análisis estadístico.

La curtosis es una medida estadística que describe la forma de la distribución de una variable aleatoria, específicamente en lo que respecta a la concentración de los datos en las colas y en el centro de la distribución. A diferencia de la asimetría, que se enfoca en la simetría de la distribución, la curtosis se focaliza en la "altura" y "anchura" de las colas de la misma.

La curtosis se consigue definir como una medida que indica el grado de apuntamiento de una distribución de probabilidad. En términos más simples, mide cuán "picuda" o "plana" es la distribución en comparación con una distribución normal. Esta medida se expresa generalmente en relación con la distribución normal, que tiene una curtosis de 3. En análisis estadísticos, a menudo se utiliza una versión ajustada de la curtosis, denominada curtosis excesiva, que se calcula restando 3 de la curtosis original. Esto posibilita que una distribución normal tenga un valor de curtosis de 0 (Moors, 1988). Existen tres tipos principales de curtosis que se utilizan para clasificar las distribuciones:

i. *Curtosis alta (leptocúrtica):* Una distribución con curtosis alta es más puntiaguda que la normal. Esto significa que tiene colas más pesadas y una mayor

concentración de datos en el centro, lo que indica una mayor probabilidad de observar valores extremos.

ii. *Curtosis baja (platicúrtica):* Una distribución con curtosis baja es más plana en comparación con la normal. Esto sugiere que los datos están más dispersos y que hay menos probabilidad de encontrar valores extremos.

iii. *Curtosis normal (mesocúrtica):* Una distribución que tiene una curtosis similar a la normal se clasifica como mesocúrtica. Este tipo de distribución es un punto de referencia en el análisis de datos.

La curtosis es fundamental para comprender la variabilidad y el comportamiento de los datos, especialmente en campos como la estadística, la economía y las ciencias sociales. Conocer la curtosis de una distribución posibilita a los analistas evaluar el riesgo y la probabilidad de eventos extremos. En finanzas, una distribución con alta curtosis consigue indicar una mayor posibilidad de pérdidas o ganancias significativas, lo que es crucial para la gestión de riesgos. En esa misma línea, la curtosis consigue influir en la selección de modelos estadísticos adecuados, ya que muchos de estos modelos asumen una distribución normal de los datos (Contento, 2019). Si la curtosis es extrema, esto consigue llevar a conclusiones erróneas si no se tiene en cuenta.

Así pues, la curtosis no solo complementa el análisis de asimetría, sino que incluso proporciona una visión más completa de la distribución de los datos, permitiendo a los investigadores y analistas tomar decisiones más informadas basadas en la estructura de los mismos. El cálculo de la asimetría y la curtosis es fundamental para poder interpretar adecuadamente la distribución de un conjunto de datos. Estas medidas nos proporcionan información valiosa sobre la forma de la distribución, lo que a su vez consigue influir en las decisiones que tomemos basadas en esos datos.

La asimetría se consigue calcular de varias maneras, pero una de las fórmulas más comunes es la asimetría de Pearson, que se define como:

$$
\text{Asimetría} = \frac{3(\text{Media} - \text{Mediana})}{\text{Desviación estándar}}
$$

Otra forma de calcular la asimetría es mediante el uso del momento de tercer orden, que se define como:

\[

\text{Asimetría} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^3

\]

donde \(n\) es el número de observaciones, \(x_i\) son los valores individuales, \(\bar{x}\) es la media, y \(s\) es la desviación estándar.

La curtosis del mismo modo se calcula a partir del momento de cuarto orden. La fórmula más utilizada es:

\[

\text{Curtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}

\]

En esta fórmula, los términos son los mismos que en la fórmula de asimetría. Es importante tener en cuenta que la curtosis se mide en relación a la distribución normal, donde una curtosis de 0 indica una distribución normal, valores positivos indican colas más pesadas (alta curtosis), y valores negativos indican colas más ligeras (baja curtosis). Para ilustrar cómo se calculan la asimetría y la curtosis, consideremos un conjunto de datos simple:

Supongamos que tenemos los siguientes valores: 2, 3, 4, 5, 6.

i. *Cálculo de la media, mediana y desviación estándar:*

- *Media (\(\bar{x}\)): \(\frac{2 + 3 + 4 + 5 + 6}{5} = 4\)*

- *Mediana: 4 (el valor del medio)*

- *Desviación estándar (\(s\)): \(\sqrt{\frac{(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2}{5}} = i.41\)*

ii. *Cálculo de la asimetría usando la fórmula de Pearson:*

\[

\text{Asimetría} = \frac{3(4 - 4)}{i.41} = 0

\]

Esto indica que la distribución es simétrica.

iii. *Cálculo de la curtosis:*

Primero, calculamos el momento de cuarto orden:

\[

\text{Curtosis} = \frac{5(5+1)}{(5-1)(5-2)(5-3)} \sum \left(\frac{x_i - 4}{i.41}\right)^4 - \frac{3(5-1)^2}{(5-2)(5-3)}

\]

Supongamos que después de calcular el momento de cuarto orden obtenemos un valor que indica una curtosis de 0, lo que sugiere que la distribución de nuestros datos es normal. Estos cálculos nos posibilitan tener una comprensión más profunda de la forma de nuestros datos y cómo se comparan con una distribución normal, lo que es crucial en el análisis estadístico.

En síntesis, la asimetría y la curtosis son dos conceptos estadísticos fundamentales que posibilitan a los analistas comprender mejor la distribución de un conjunto de datos. La asimetría proporciona información valiosa sobre la simetría de la distribución, indicando si los datos tienden a agruparse más hacia un lado que hacia el otro (Martínez, 2009). Por otro lado, la curtosis presenta una visión sobre la "altura" y la "anchura" de la distribución, revelando la presencia de valores extremos y la concentración de datos en torno a la media.

La correcta interpretación de la asimetría y la curtosis no solo es crucial para la descripción de los datos, sino que al igual influye en la selección de métodos estadísticos adecuados para el análisis. En situaciones donde se detecta una alta asimetría, consigue ser necesario aplicar transformaciones a los datos para cumplir con las suposiciones de normalidad requeridas por ciertos métodos estadísticos. Asimismo, al observar una curtosis alta, los analistas consiguen estar alerta ante la posibilidad de outliers que podrían afectar sus conclusiones.

Por tanto, el cálculo e interpretación de la asimetría y la curtosis son herramientas poderosas que posibilitan a los investigadores y profesionales de diversas disciplinas tomar decisiones informadas basadas en la naturaleza de sus datos. Al integrar estos conceptos en el análisis estadístico, se mejora

significativamente la calidad de las inferencias y se facilita una comprensión más profunda de los fenómenos que se estudian. Dominar la asimetría y la curtosis es esencial para cualquier estadístico o investigador que busque obtener una visión clara y precisa de sus datos.

1.3 Cómo Calcular y Aplicar el Coeficiente de Correlación Biserial entre Variables Cuantitativas y Binarias

El análisis de datos es una herramienta fundamental en la investigación, permitiendo a científicos y académicos extraer conclusiones significativas a partir de información cuantitativa. En este contexto, la correlación entre variables juega un papel crucial, ya que posibilita identificar y cuantificar la relación que existe entre distintas medidas. En particular, el coeficiente de correlación biserial se presenta como un método estadístico valioso para explorar la relación entre una variable cuantitativa y una variable binaria.

Antes de adentrarnos en el cálculo del coeficiente de correlación biserial, es fundamental comprender algunos conceptos básicos que nos ayudarán a contextualizar su uso y aplicación. En este apartado, abordaremos las definiciones de variable cuantitativa y variable binaria, así como la importancia de la correlación en el ámbito de la estadística. Las variables cuantitativas son aquellas que se consiguen medir y expresar numéricamente. Estas variables posibilitan realizar operaciones matemáticas y se dividen en dos categorías principales: variables discretas y variables continuas.

- *Variables discretas:* Toman valores específicos y contables, como el número de hijos en una familia o el número de respuestas correctas en un examen.

- *Variables continuas:* Consiguen tomar cualquier valor dentro de un rango determinado, como la altura, el peso o la temperatura.

La naturaleza numérica de las variables cuantitativas facilita el análisis estadístico, permitiendo a los investigadores identificar patrones, tendencias y relaciones entre diferentes conjuntos de datos. Las variables binarias, también conocidas como variables dicotómicas, son aquellas que solo consiguen tomar dos posibles valores. Estos valores suelen representarse como 0 y 1, donde cada uno indica una categoría o estado distinto. En un estudio sobre la presencia o ausencia de una enfermedad, los valores podrían ser "1" para indicar la presencia y "0" para la ausencia.

El uso de variables binarias es común en diversos campos de investigación, ya que posibilitan simplificar la complejidad de fenómenos que, aunque consiguen ser multifacéticos, se consiguen clasificar en dos categorías principales. Esto facilita la comparación y el análisis de datos, especialmente cuando se combinan con variables cuantitativas.

La correlación es un concepto central en estadística que se refiere a la relación entre dos o más variables. Un coeficiente de correlación mide la fuerza y la dirección de esta relación. En el contexto de la investigación, comprender la correlación entre variables es crucial, ya que posibilita a los investigadores identificar patrones y hacer inferencias sobre cómo una variable consigue influir en otra (Rodríguez et al., 2021). La correlación consigue ser positiva, negativa o nula. Una correlación positiva indica que, a medida que una variable aumenta, la otra al igual tiende a aumentar. En contraste, una correlación negativa implica que, a medida que una variable aumenta, la otra tiende a disminuir. La correlación nula sugiere que no hay relación alguna entre las variables.

Entender estos conceptos es esencial para aplicar correctamente el coeficiente de correlación biserial, que nos permitirá explorar la relación entre una variable cuantitativa y una variable binaria. El coeficiente de correlación biserial es una herramienta estadística utilizada para medir la relación entre una variable cuantitativa y una variable binaria. Este coeficiente es especialmente útil en contextos donde se desea evaluar cómo una variable dicotómica influye o se relaciona con una variable continua. El coeficiente de correlación biserial se denota comúnmente como r_{bis} y se calcula utilizando la siguiente fórmula:

$$
r_{bis} = \frac{(M_1 - M_0)}{S} \cdot \sqrt{\frac{p(1-p)}{n}}
$$

donde:

- M_1 es la media de la variable cuantitativa para el grupo que presenta el valor 1 en la variable binaria.

- M_0 es la media de la variable cuantitativa para el grupo que presenta el valor 0 en la variable binaria.

- S es la desviación estándar de la variable cuantitativa.

- p es la proporción de casos que tienen el valor 1 en la variable binaria.

- n es el tamaño total de la muestra.

Esta fórmula posibilita calcular el coeficiente considerando la diferencia de medias de la variable cuantitativa entre los dos grupos definidos por la variable binaria, ajustada por la variabilidad de los datos. Imaginemos un estudio donde se analiza la relación entre el rendimiento académico (variable cuantitativa) y el hecho de haber recibido tutorías (variable binaria: 1 = sí, 0 = no) de un grupo de 30 estudiantes. Supongamos que los resultados son los siguientes:

- *Media de los estudiantes que recibieron tutorías (M_1):* 85

- *Media de los estudiantes que no recibieron tutorías (M_0):* 75

- *Desviación estándar (S):* 10

- *Proporción de estudiantes que recibieron tutorías (p):* 0.6 (18 estudiantes)

- *Tamaño total de la muestra (n):* 30

Sustituyendo estos valores en la fórmula:

$$
r_{bis} = \frac{(85 - 75)}{10} \cdot \sqrt{\frac{0.6(1-0.6)}{30}}
$$

$$
r_{bis} = \frac{10}{10} \cdot \sqrt{\frac{0.6 \cdot 0.4}{30}}
$$

$$
r_{bis} = 1 \cdot \sqrt{\frac{0.24}{30}}
$$

$$
r_{bis} = 1 \cdot \sqrt{0.008}
$$

$$
r_{bis} \approx 1 \cdot 0.0894 \approx 0.0894
$$

Esto indica que existe una correlación positiva baja entre haber recibido tutorías y el rendimiento académico, lo que sugiere que, en promedio, aquellos que recibieron apoyo adicional tienden a tener un mejor rendimiento. El coeficiente de correlación biserial, como cualquier indicador de correlación, oscila entre -1 y i. Un valor de r_{bis} de 0 indica que no hay correlación entre las variables, mientras que valores cercanos a 1 o -1 indican una correlación positiva o negativa fuerte, respectivamente. En nuestro ejemplo, un r_{bis} de aproximadamente 0.0894 sugiere una correlación positiva débil, lo que implica que existe una tendencia leve en la que los estudiantes que recibieron tutorías tienden a tener mejores calificaciones.

Es importante recordar que la correlación no implica causalidad. Sin obstar los resultados sugieren una relación, no se consigue concluir directamente que las tutorías son la causa del mejor rendimiento académico sin considerar otros factores que puedan estar influyendo en los resultados. El coeficiente de correlación biserial es una herramienta estadística valiosa que encuentra aplicación en diversas disciplinas. Su capacidad para medir la relación entre una variable cuantitativa y una variable binaria lo convierte en un recurso esencial en muchos campos de investigación.

En el ámbito de la psicología, el coeficiente biserial se utiliza frecuentemente para evaluar la relación entre variables que consiguen influir en el comportamiento humano. Los investigadores consiguen estar interesados en analizar cómo el nivel de ansiedad (una variable cuantitativa) se relaciona con la presencia o ausencia de un trastorno mental (una variable binaria) (Menses et al., 2013). Al calcular el coeficiente biserial en este contexto, los psicólogos consiguen obtener información valiosa sobre la intensidad de la relación entre estas variables, lo que consigue ayudar en el diseño de tratamientos o intervenciones.

En el campo de la medicina, el coeficiente de correlación biserial consigue ser fundamental para establecer relaciones entre variables clínicas. Se consigue utilizar para investigar cómo diferentes niveles de colesterol (variable cuantitativa) se correlacionan con el riesgo de desarrollar una enfermedad

cardíaca (variable binaria: sí/no). Este tipo de análisis proporciona a médicos e investigadores información crítica que consigue influir en las decisiones de diagnóstico y tratamiento, así como en la formulación de recomendaciones de salud pública.

El coeficiente biserial también es ampliamente utilizado en el análisis de encuestas, donde se busca entender la relación entre respuestas cuantitativas y características demográficas binarias. En una encuesta sobre hábitos de consumo, los investigadores consiguen querer examinar cómo el ingreso mensual (variable cuantitativa) se relaciona con la decisión de comprar un producto específico (variable binaria: sí/no). Al aplicar el coeficiente biserial, se consiguen identificar patrones que informen estrategias de marketing y segmentación de mercado.

En definitiva, el coeficiente de correlación biserial es una herramienta versátil que posibilita a investigadores y profesionales de diversas disciplinas explorar y comprender mejor las relaciones entre variables cuantitativas y binarias. Su aplicación en psicología, medicina y análisis de encuestas demuestra su importancia en la obtención de conclusiones significativas y la toma de decisiones basadas en datos.

El cálculo del coeficiente de correlación biserial entre una variable cuantitativa y una variable binaria es una herramienta estadística valiosa que posibilita explorar y entender la relación entre distintos tipos de datos. La correlación biserial no solo proporciona una medida cuantitativa de la relación existente entre las variables, sino que siempre ofrece una visión de cómo las diferencias en una variable cuantitativa consiguen estar asociadas con las categorías de una variable binaria (Mias y Tornimbeni, 2020). Esto es especialmente relevante en campos como la psicología, las ciencias de la salud y el análisis de encuestas, donde las decisiones informadas y las conclusiones dependen de la comprensión precisa de las relaciones entre variables.

Al aplicar este coeficiente, los investigadores consiguen obtener percepciones significativas que consiguen influir en la toma de decisiones y el diseño de futuras investigaciones. La capacidad de interpretar correctamente los resultados del coeficiente biserial es crucial, ya que posibilita a analistas y científicos sociales comunicar de manera efectiva sus hallazgos y su relevancia en contextos prácticos. En otras palabras, el coeficiente de correlación biserial no solo es un recurso técnico, sino al igual un puente hacia una mejor comprensión

de las dinámicas complejas que existen entre las variables que estudiamos. Al dominar esta herramienta, los investigadores consiguen dignificar su análisis y contribuir al avance del conocimiento en sus respectivos campos.

1.4 Pruebas de Normalidad: Realización e Interpretación de la Prueba de Shapiro-Wilk y Otras Alternativas Estadísticas

La normalidad de los datos es un supuesto fundamental en muchas técnicas estadísticas. La capacidad de aplicar inferencias y modelos estadísticos con confianza depende en gran medida de si los datos siguen una distribución normal. En este sentido, las pruebas de normalidad se han convertido en herramientas esenciales para los estadísticos y analistas de datos, ya que posibilitan evaluar si un conjunto de datos se ajusta a la distribución normal.

La normalidad es una característica deseable en los datos porque muchas pruebas estadísticas, como la prueba t de Student y el análisis de la varianza (ANOVA), asumen que los datos son distribuidos normalmente. Cuando esta suposición se viola, los resultados de estas pruebas consiguen ser engañosos y llevar a conclusiones incorrectas (Sánchez et al., 2024). Por lo tanto, verificar la normalidad de los datos es un paso crítico en el análisis estadístico. La normalidad no solo afecta la validez de las pruebas inferenciales, sino que del mismo modo impacta en la calidad de los intervalos de confianza y en las estimaciones de los parámetros.

Las pruebas de normalidad son especialmente útiles en diversas áreas de investigación, incluyendo las ciencias sociales, la biología, la economía y la ingeniería. En estudios clínicos, es fundamental asegurarse de que las mediciones (como la presión arterial o los niveles de colesterol) se distribuyan normalmente para aplicar adecuadamente pruebas estadísticas que evalúen la efectividad de un tratamiento. Ahora bien, en el análisis de calidad en manufactura, la normalidad de los datos de producción consigue influir en la toma de decisiones sobre procesos y mejoras.

Entre las pruebas de normalidad más utilizadas se encuentra la prueba de Shapiro-Wilk, desarrollada en 1965. Esta prueba es especialmente conocida por su alta potencia en la detección de desviaciones de la normalidad, incluso con tamaños de muestra pequeños. La prueba de Shapiro-Wilk compara la distribución de los datos observados con una distribución normal y proporciona

un estadístico que posibilita decidir si se rechaza o no la hipótesis nula de normalidad.

La prueba de Shapiro-Wilk es una de las más utilizadas para evaluar la normalidad de un conjunto de datos. Su método de cálculo se basa en la comparación de la distribución de los datos observados con la distribución normal. Para llevar a cabo esta prueba, se siguen los siguientes pasos:

i. *Recopilación de datos:* Se inicia con un conjunto de datos que se desea evaluar. Es importante que la muestra sea representativa y que contenga al menos 3 observaciones, si bien lo ideal es contar con un tamaño de muestra mayor a 30 para obtener resultados más robustos.

ii. *Cálculo de estadísticas:* Se calcula la media y la varianza de los datos. Luego, se ordenan los datos de menor a mayor y se asignan coeficientes a cada uno de ellos, los cuales dependen de la varianza de la muestra y de la media.

iii. *Cálculo del estadístico W:* El estadístico W se calcula como la razón entre la suma de los cuadrados de las diferencias de los datos ordenados respecto a la media y la suma de los cuadrados de las diferencias de los datos respecto a la media. Un valor de W cercano a 1 indica que los datos son normalmente distribuidos.

iv. *Valor p y decisión:* Se compara el valor de W obtenido con los valores críticos de la distribución de Shapiro-Wilk o se calcula el valor p asociado. Si el valor p es menor que el nivel de significancia (comúnmente 0.05), se rechaza la hipótesis nula de que los datos provienen de una distribución normal.

La interpretación de los resultados de la prueba de Shapiro-Wilk se canaliza en el valor del estadístico W y el valor p.

- Si el estadístico W es alto y el valor p es mayor que el nivel de significancia, no hay evidencia suficiente para rechazar la hipótesis nula, lo que sugiere que los datos siguen una distribución normal.

- En cambio, si el estadístico W es bajo y el valor p es menor que el nivel de significancia, se rechaza la hipótesis nula, indicando que los datos no se distribuyen normalmente.

Es esencial considerar que la prueba de Shapiro-Wilk es sensible al tamaño de la muestra. En muestras pequeñas, consigue no detectar desviaciones de la normalidad, mientras que en muestras grandes, incluso pequeñas desviaciones consiguen resultar en un valor p significativo. Por lo tanto, es recomendable

complementar esta prueba con métodos gráficos, como histogramas o gráficos Q-Q, para obtener una evaluación más completa de la normalidad de los datos. Para ilustrar la aplicación de la prueba de Shapiro-Wilk, consideremos dos ejemplos:

i. *Ejemplo 1. Datos de altura:* Supongamos que se tiene una muestra de 50 individuos y se mide su altura en centímetros. Al aplicar la prueba de Shapiro-Wilk, se obtiene un estadístico W de 0.97 y un valor p de 0.3iv. Dado que el valor p es mayor que 0.05, podemos concluir que no hay evidencia suficiente para rechazar la hipótesis de normalidad en la distribución de las alturas.

ii. *Ejemplo 2. Datos de puntuaciones de un examen:* En otro caso, se recopilan las puntuaciones de un examen de 30 estudiantes y se aplica la prueba de Shapiro-Wilk. Se obtiene un estadístico W de 0.85 y un valor p de 0.0i. Aquí, el valor p es menor que 0.05, por lo que rechazamos la hipótesis nula, sugiriendo que las puntuaciones no se distribuyen normalmente.

Estos ejemplos destacan la utilidad de la prueba de Shapiro-Wilk en diferentes contextos y la importancia de interpretar sus resultados en función del tamaño de la muestra y del contexto del análisis. Además de la prueba de Shapiro-Wilk, existen varias otras pruebas que los estadísticos utilizan para evaluar la normalidad de un conjunto de datos. Cada una de estas pruebas tiene sus propias características, ventajas y desventajas.

La prueba de Kolmogorov-Smirnov (K-S) es una de las pruebas más antiguas y ampliamente utilizadas para evaluar la normalidad de los datos. Esta prueba se basa en la comparación de la función de distribución empírica de la muestra con la función de distribución acumulativa de la distribución normal teórica (Luzuriaga et al., 2023) El procedimiento implica calcular la máxima diferencia absoluta entre las dos funciones de distribución. Si esta diferencia excede un umbral crítico, se rechaza la hipótesis nula de que los datos siguen una distribución normal. Entre las ventajas de la prueba K-S es que consigue aplicarse a muestras de cualquier tamaño, pero su eficacia disminuye en muestras pequeñas o en distribuciones que no son perfectamente normales.

La prueba de Anderson-Darling es otra alternativa popular para evaluar la normalidad. A diferencia de la prueba K-S, que se basa en la máxima diferencia, la prueba de Anderson-Darling toma en cuenta la forma de la distribución y otorga más peso a las colas de la distribución. Esto la hace

especialmente útil cuando se desea detectar desviaciones de la normalidad en las partes extremas de la distribución.

La estadística de prueba se calcula a partir de la suma de los cuadrados de las diferencias entre la función de distribución empírica y la función de distribución normal teórica, ponderada por la varianza. Al igual que la prueba K-S, si el valor calculado supera un cierto umbral crítico, se rechaza la hipótesis nula de normalidad. La prueba de Anderson-Darling es generalmente considerada más potente que la K-S, especialmente para muestras más pequeñas.

Al comparar las pruebas de normalidad, es importante considerar varios factores, como el tamaño de la muestra, la sensibilidad a las colas y la naturaleza de los datos. La prueba de Shapiro-Wilk es a menudo preferida para muestras pequeñas debido a su alta potencia, mientras que la K-S se utiliza para conjuntos de datos más grandes y se consigue aplicar a distribuciones diferentes a la normal (Roco et al., 2023). Por otro lado, la prueba de Anderson-Darling se destaca por su capacidad para detectar desviaciones en las colas de la distribución.

En suma, si bien la prueba de Shapiro-Wilk es una herramienta valiosa para la evaluación de la normalidad, es fundamental considerar otras pruebas como K-S y Anderson-Darling para obtener una comprensión más completa de la distribución de los datos. La elección de la prueba adecuada dependerá del contexto específico del análisis y de las características de los datos en cuestión. La normalidad de los datos es un supuesto fundamental en muchas técnicas estadísticas, como la regresión y los tests paramétricos. La prueba de Shapiro-Wilk se ha destacado por su eficacia y su capacidad para detectar desviaciones de la normalidad, incluso en muestras pequeñas. Entonces, hemos discutido otras pruebas de normalidad, como la de Kolmogorov-Smirnov y la de Anderson-Darling, que brindan alternativas útiles dependiendo del contexto y del tamaño de la muestra.

Elegir la prueba de normalidad correcta es crucial, ya que no todas las pruebas son igualmente efectivas en todas las situaciones. La prueba de Shapiro-Wilk es generalmente preferida para muestras pequeñas, mientras que la prueba de Kolmogorov-Smirnov consigue ser más adecuada para tamaños de muestra más grandes. La prueba de Anderson-Darling, por su parte, ofrece una sensibilidad mejorada en las colas de la distribución, lo que consigue ser relevante en ciertos análisis. Por lo tanto, es fundamental considerar el tamaño

de la muestra, la naturaleza de los datos y el objetivo del análisis al seleccionar la prueba más adecuada.

Para investigadores y analistas, se recomienda llevar a cabo una evaluación exhaustiva de la normalidad de los datos antes de aplicar técnicas estadísticas que asuman normalidad. Además, se sugiere complementar los resultados de las pruebas de normalidad con representaciones gráficas, como histogramas y diagramas de caja, para obtener una visión más completa de la distribución de los datos. Se alienta a la comunidad científica a continuar investigando y desarrollando nuevas pruebas de normalidad y métodos de evaluación que puedan mejorar la precisión y la robustez del análisis estadístico. Esto no solo enaltecerá el cuerpo de conocimiento existente, sino que del mismo modo facilitará la interpretación y el uso de los resultados en diversas disciplinas.

Capítulo II

Cuantiles, percentiles e intervalos de confianza: Remuestreos bootstrap

2.1 Cuantiles y Percentiles: Cálculo y Aplicaciones en Intervalos de Confianza

Los cuantiles y percentiles son conceptos fundamentales en el ámbito de la estadística, utilizados para describir la distribución de un conjunto de datos. A través de estos conceptos, se consigue obtener una visión más clara de cómo se distribuyen los valores en una población o muestra, permitiendo así realizar inferencias significativas.

Los cuantiles son puntos de corte que dividen un conjunto de datos en segmentos iguales. El cuartil divide los datos en cuatro partes iguales, mientras que el quintil lo hace en cinco. En este contexto, los percentiles son un tipo específico de cuantiles que dividen los datos en cien partes iguales. Es decir, el percentil 25 (P25) representa el valor bajo el cual se encuentra el 25% de los datos. De esta manera, tanto los cuantiles como los percentiles posibilitan identificar posiciones relativas dentro de una distribución, facilitando la comprensión de la variabilidad y la tendencia central de los datos.

La utilización de cuantiles y percentiles es crucial en la estadística descriptiva, ya que presentan una forma efectiva de resumir y representar grandes volúmenes de información. Estos indicadores no solo proporcionan información sobre la localización de los datos, sino que también son herramientas valiosas para identificar outliers o valores atípicos, así como para comparar diferentes conjuntos de datos. En suma, en el contexto de la inferencia estadística, los cuantiles y percentiles son esenciales para la construcción de intervalos de confianza, ya que posibilitan establecer rangos dentro de los cuales se espera que se encuentren ciertos parámetros poblacionales.

Los intervalos de confianza son estimaciones que indican la incertidumbre sobre un parámetro poblacional. Los cuantiles y percentiles juegan un papel fundamental en la determinación de estos intervalos. Al calcular un intervalo de

confianza del 95% para la media de una población, se utilizan los percentiles ii.5 y 97.5 de la distribución de la muestra para definir los límites inferior y superior del intervalo. De esta forma, los cuantiles y percentiles no solo ayudan a describir la distribución de los datos, sino que al igual son herramientas clave en la inferencia estadística, proporcionando un marco para realizar afirmaciones sobre la población a partir de muestras. El método de ordenación es uno de los planteamientos más sencillos para calcular cuantiles. Consiste en seguir estos pasos:

i. *Recolección de datos:* Reúne todos los datos que deseas analizar.

ii. *Ordenar los datos:* Organiza los datos en orden ascendente. Esta etapa es crucial, ya que los cuantiles dependen de la posición relativa de los valores en el conjunto de datos.

iii. *Cálculo del cuantil:* Para encontrar un cuantil específico, como el percentil 25 (Q1), se utiliza la fórmula:

\[

P_k = \frac{k(n + 1)}{100}

\]

donde \(P_k \) es la posición del percentil \(k \) y \(n \) es el número total de observaciones. Se redondea al entero más cercano para identificar la posición del dato en la lista ordenada.

iv. *Extracción del valor:* El cuantil se obtiene directamente del conjunto de datos ordenados, seleccionando el valor que se encuentra en la posición calculada.

Existen fórmulas más sofisticadas que posibilitan calcular cuantiles sin necesidad de ordenar los datos. Estas fórmulas son útiles en situaciones donde se trabaja con grandes volúmenes de datos o distribuciones complejas. Un ejemplo común es el uso de la interpolación para estimar cuantiles en conjuntos de datos que no se distribuyen uniformemente. Para calcular el percentil \(k \) (el percentil 90), la fórmula general consigue ser:

\[

P_k = \text{Valor en la posición } \left(\frac{k(n - 1)}{100} + 1\right)

\]

En este caso, si la posición calculada no es un número entero, se utiliza la interpolación entre los dos valores más cercanos en el conjunto de datos. Estas fórmulas son particularmente útiles en el análisis teórico y en situaciones donde se requiere un tratamiento más analítico. Hoy en día, el uso de software estadístico ha facilitado enormemente el cálculo de cuantiles. Herramientas como R, Python (con bibliotecas como NumPy y Pandas), SPSS, y Excel dedican funciones integradas que posibilitan calcular cuantiles de manera rápida y eficiente.

i. *R:* Utiliza la función quantile(), donde consigues especificar el conjunto de datos y el percentil deseado.

ii. *Python:* Con la biblioteca NumPy, el comando numpy.percentile(data, k) devuelve el percentil \(k \) del conjunto de datos.

iii. *SPSS:* Proporciona opciones en su menú de análisis descriptivo para calcular percentiles.

iv. *Excel:* Usa la función PERCENTILE() o PERCENTILE.INC() para obtener percentiles directamente desde una hoja de cálculo.

Estas herramientas no solo simplifican el proceso, sino que al igual minimizan el riesgo de errores en cálculos manuales, permitiendo a los analistas centrarse en la interpretación de los resultados en lugar de en la mecánica del cálculo. En síntesis, existen diversos métodos para calcular cuantiles, cada uno con sus ventajas y desventajas. La elección del método dependerá del contexto y la naturaleza de los datos a analizar. Los percentiles son herramientas cruciales en el análisis estadístico, sobre todo cuando se trata de la construcción de intervalos de confianza.

Cuando se trabaja con muestras de datos, los percentiles consiguen proporcionar estimaciones robustas de la media y la mediana, así como de otros parámetros relevantes. El percentil 50 representa la mediana, que es un estimador menos sensible a los valores atípicos en comparación con la media aritmética. Al construir intervalos de confianza, se consiguen utilizar percentiles para determinar el rango de valores en el que se espera que se encuentre el parámetro de interés con un nivel de confianza específico (Batanero y Díaz, 2011).

En la práctica, si deseamos calcular un intervalo de confianza para la media de una población, podemos recurrir a la distribución muestral de la media

y utilizar percentiles de esta distribución para establecer los límites del intervalo de confianza. Este planteamiento es especialmente útil en situaciones donde la distribución de los datos no es normal, permitiendo así una estimación más precisa. La interpretación de los resultados obtenidos a partir de los percentiles en intervalos de confianza es fundamental para la toma de decisiones. Cuando se comunica el intervalo de confianza a un público no especializado, es vital explicitar que este intervalo ofrece un rango de valores plausibles para el parámetro estimado, basándose en los datos de la muestra analizada.

Si un intervalo de confianza del 95% para la media de una muestra está entre 10 y 20, esto implica que, si se repitiera el muestreo un gran número de veces, aproximadamente el 95% de los intervalos calculados a partir de esas muestras incluirían el verdadero valor de la media de la población. Esta interpretación ayuda a los investigadores y tomadores de decisiones a comprender la incertidumbre inherente a sus estimaciones y a evaluar el riesgo en sus decisiones basadas en datos. Para ilustrar el uso de percentiles en intervalos de confianza, consideremos un ejemplo práctico: supongamos que un investigador está interesado en estimar la duración promedio de una visita a un parque de atracciones. Después de realizar una encuesta a 100 visitantes, obtiene los siguientes tiempos de permanencia en minutos: [45, 50, 55, 60, 65, 70, 75, 80, 85, 90].

Para calcular un intervalo de confianza del 95% para la media, el investigador podría calcular el percentil ii.5 y el percentil 97.5 de la muestra. Estos percentiles proporcionan los límites del intervalo de confianza, que da una idea de la variabilidad y la incertidumbre en las estimaciones de tiempo de visita. Si, por ejemplo, el percentil ii.5 resulta ser 54 minutos y el percentil 97.5 es 82 minutos, el investigador consigue concluir que tiene un 95% de confianza en que la duración promedio de la visita al parque se encuentra entre esos valores. Dicho de otro modo, los percentiles no solo facilitan la estimación de parámetros estadísticos, sino que siempre ennoblecen la interpretación de los resultados y ofrecen ejemplos concretos que demuestran su utilidad en la práctica estadística.

En este capítulo, hemos explorado la importancia de los cuantiles y percentiles en el ámbito de la estadística, así como su relación directa con los intervalos de confianza. Comenzamos definiendo estos conceptos fundamentales, destacando cómo los cuantiles dividen un conjunto de datos en

partes iguales y los percentiles posibilitan identificar la posición relativa de un valor en un conjunto. Luego, revisamos los métodos utilizados para calcular cuantiles, incluyendo el método de ordenación, las fórmulas estadísticas pertinentes y las herramientas de software que facilitan estos cálculos. Ahora bien, discutimos las aplicaciones de los percentiles en la estimación de parámetros y la interpretación de resultados en el contexto de intervalos de confianza, proporcionando ejemplos prácticos que ilustran su utilidad en investigaciones y análisis de datos.

Los cuantiles y percentiles son herramientas poderosas que posibilitan a los estadísticos y analistas tomar decisiones informadas basadas en datos. Su capacidad para presentar una comprensión clara de la distribución de los datos y su relación con los intervalos de confianza los convierte en elementos esenciales en el análisis estadístico (Posada, 2016). En este sentido, la correcta interpretación y aplicación de cuantiles y percentiles no solo mejora la calidad de los resultados estadísticos, sino que también potencia la confianza en las decisiones que se derivan de ellos. La estadística, al fin y al cabo, no es solo sobre números, sino sobre la historia que estos números cuentan, y los cuantiles y percentiles juegan un papel central en esa narrativa.

2.2 Remuestreos bootstrap, desviación típica e intervalos de confianza

El remuestreo bootstrap es una técnica estadística poderosa que ha ganado popularidad en los últimos años debido a su capacidad para proporcionar estimaciones robustas y confiables en situaciones donde los métodos tradicionales consiguen ser inadecuados. Esta metodología se basa en la idea de que, a partir de un conjunto de datos muestrales, es posible realizar múltiples réplicas de la muestra original para obtener una mejor comprensión de la variabilidad y la incertidumbre asociada a las estimaciones estadísticas.

El término "bootstrap" proviene de la expresión en inglés "to pull oneself up by one's bootstraps", que implica mejorar una situación sin ayuda externa. En el contexto estadístico, el remuestreo bootstrap consiste en generar múltiples muestras de datos (denominadas "muestras bootstrap") al seleccionar aleatoriamente con reemplazo de la muestra original. Esto significa que algunos datos consiguen incluirse varias veces en una muestra bootstrap, mientras que otros consiguen no ser seleccionados en absoluto. A través de este proceso, se

consigue obtener una distribución empírica de estimaciones que posibilita realizar inferencias sobre la población a partir de la cual se extrajo la muestra original.

La técnica de remuestreo bootstrap es especialmente valiosa en el análisis estadístico porque proporciona una forma de calcular intervalos de confianza y estimaciones de error estándar sin asumir que los datos siguen una distribución específica, como la normal. Esto la convierte en una herramienta esencial en situaciones donde se tienen muestras pequeñas o donde los datos presentan características que violan las suposiciones de los métodos paramétricos clásicos. Por añadidura, el bootstrap posibilita a los investigadores evaluar la estabilidad de sus estimaciones y obtener una mejor comprensión de las variaciones inherentes en sus datos.

La desviación típica, también conocida como desviación estándar, es una medida que cuantifica la dispersión o variabilidad de un conjunto de datos. Se utiliza para determinar en qué medida los valores individuales de una distribución se alejan de la media (Quevedo, 2011). Al ser una medida de dispersión, proporciona información crucial acerca de la distribución de los datos: una desviación típica baja indica que los datos están más concentrados alrededor de la media, mientras que una desviación típica alta sugiere que los datos están más dispersos.

Matemáticamente, la desviación típica se define como la raíz cuadrada de la varianza. La varianza, a su vez, es el promedio de las diferencias al cuadrado entre cada dato y la media. La fórmula para calcular la desviación típica σ de una población es:

$$
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}
$$

donde N es el tamaño de la población, x_i son los valores individuales y μ es la media de la población. En el caso de una muestra, se utiliza un ajuste en el denominador para obtener la desviación típica muestral s:

$$
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}

\]

donde n es el tamaño de la muestra y $\bar{x}$ es la media muestral.

El cálculo de la desviación típica en una muestra es fundamental en la estadística, ya que posibilita estimar la variabilidad de los datos recolectados. Para llevar a cabo este cálculo, los pasos que se deben seguir son los siguientes:

i. *Recopilación de datos:* Se debe tener un conjunto de datos que represente la muestra de interés.

ii. *Cálculo de la media muestral:* Se suman todos los valores de la muestra y se divide por el número de observaciones.

iii. *Cálculo de las diferencias al cuadrado:* Para cada valor de la muestra, se resta la media muestral y se eleva al cuadrado.

iv. *Promedio de las diferencias al cuadrado:* Se suman todas las diferencias al cuadrado y se dividen por $n-1$ (donde n es el tamaño de la muestra) para obtener la varianza.

5. *Raíz cuadrada de la varianza:* Se toma la raíz cuadrada de la varianza para obtener la desviación típica.

Este proceso posibilita a los investigadores entender mejor la variabilidad de los datos y, por ende, realizar inferencias más precisas sobre la población de la que se extrajo la muestra. La interpretación de la desviación típica es esencial para el análisis de datos en diversos campos, desde la investigación científica hasta las finanzas. En contextos prácticos, la desviación típica ayuda a los estadísticos y analistas a responder preguntas como: ¿qué tan consistente es el rendimiento de un estudiante en relación a sus compañeros? o ¿cuánto varía el retorno de una inversión en el mercado de valores?

En un estudio educativo, una baja desviación típica en las calificaciones de los estudiantes indicaría que la mayoría de ellos obtuvo resultados similares, sugiriendo una enseñanza uniforme. En contraste, una alta desviación típica podría señalar diferencias significativas en el rendimiento, lo que podría requerir atención adicional. En el ámbito financiero, la desviación típica se utiliza para medir el riesgo de una inversión. Una inversión con una alta desviación típica es considerada más arriesgada, ya que su rendimiento consigue fluctuar

significativamente, mientras que una inversión con una baja desviación típica es más estable y predecible.

En definitiva, la desviación típica no solo es una herramienta matemática, sino que del mismo modo ofrece un marco práctico para entender la variabilidad en diferentes contextos, permitiendo a los investigadores y profesionales tomar decisiones informadas basadas en los datos. Los intervalos de confianza son una herramienta estadística fundamental que posibilita estimar un rango de valores en el cual se espera que se encuentre un parámetro poblacional desconocido, con un nivel de confianza específico.

En términos simples, un intervalo de confianza proporciona una estimación de la incertidumbre asociada a una medida calculada a partir de una muestra. Si se calcula un intervalo de confianza del 95% para la media de una población, esto implica que si se repitieran múltiples muestras y se calcularan intervalos de confianza para cada una, aproximadamente el 95% de esos intervalos incluirían la verdadera media poblacional (Dagnino, 2014). Existen varios métodos para calcular intervalos de confianza, siendo los más comunes:

i. *Método paramétrico:* Este método asume que los datos siguen una distribución normal. Para construir un intervalo de confianza para la media, se utiliza la fórmula:

$$
\bar{x} \pm z \left(\frac{s}{\sqrt{n}}\right)
$$

donde $\bar{x}$ es la media muestral, z es el valor crítico de la distribución normal para el nivel de confianza deseado, s es la desviación típica de la muestra, y n es el tamaño de la muestra.

ii. *Método no paramétrico:* Cuando no se consigue asumir la normalidad de los datos, se consiguen utilizar métodos no paramétricos, como el método de percentiles o el método bootstrap. En el planteamiento bootstrap, se generan múltiples muestras con reemplazo de la muestra original, y se calcula el estimador de interés (la media) para cada una de ellas. Luego, se utiliza la distribución de esos estimadores para construir el intervalo de confianza.

iii. *Método basado en la t de Student:* Para muestras pequeñas (generalmente \(n < 30\)), se utiliza la distribución t de Student en lugar de la distribución normal. Esto se debe a que la estimación de la desviación típica es menos precisa en muestras pequeñas, y la distribución t ajusta mejor la incertidumbre.

Los intervalos de confianza son ampliamente utilizados en diversas áreas de investigación, incluyendo medicina, psicología, ciencias sociales y economía. Su aplicación posibilita a los investigadores:

i. *Evaluar la precisión de las estimaciones:* Proporcionan un contexto para interpretar los resultados, permitiendo a los investigadores y tomadores de decisiones entender la incertidumbre asociada a las estimaciones obtenidas.

ii. *Comparar grupos:* Al calcular intervalos de confianza para diferentes grupos, los investigadores consiguen determinar si hay diferencias significativas entre ellos. Si los intervalos de confianza no se superponen, esto consigue indicar una diferencia estadísticamente significativa.

iii. *Informar políticas y prácticas:* En el ámbito de la salud pública, los intervalos de confianza ayudan a informar decisiones sobre intervenciones y políticas basadas en la evidencia, al proporcionar una estimación clara de la efectividad de un tratamiento o programa.

En palabras, los intervalos de confianza son una herramienta esencial en el análisis estadístico que no solo posibilita estimar parámetros poblacionales, sino que al igual ayuda a comunicar la incertidumbre de manera efectiva, desempeñando un papel crucial en la interpretación y aplicación de los resultados de investigación.

El uso de remuestreos bootstrap ha revolucionado la forma en que los estadísticos abordan problemas de estimación y validación. Al proporcionar una manera de calcular intervalos de confianza y estimaciones de error sin depender de distribuciones paramétricas, el bootstrap se convierte en una herramienta invaluable, especialmente en situaciones donde los datos son limitados o no se ajustan a las suposiciones tradicionales. A través de esta técnica, los investigadores consiguen obtener resultados más confiables y realistas, lo que a su vez potencia la calidad de la toma de decisiones basadas en datos.

Además, el bootstrap no solo se limita a la estimación de intervalos de confianza; su aplicabilidad se extiende a diversas áreas de la estadística,

incluyendo la regresión, clasificación y análisis de supervivencia. Esta versatilidad resalta la importancia del remuestreo bootstrap en el contexto actual de la estadística, donde la complejidad de los datos y la necesidad de análisis más profundos son cada vez más comunes.

A medida que la estadística continúa evolucionando, el remuestreo bootstrap seguirá siendo un área de interés y desarrollo. Las futuras investigaciones podrían enfocarse en mejorar los algoritmos existentes para aumentar la eficiencia del remuestreo, así como en la integración del bootstrap con técnicas de aprendizaje automático. En espacial, la exploración de métodos bootstrap en datos de alta dimensión y en contextos no paramétricos podría ofrecer nuevas perspectivas y desafíos para los estadísticos.

Asimismo, la educación y la divulgación sobre el uso del remuestreo bootstrap son esenciales. A medida que los investigadores y analistas de datos se familiarizan con esta herramienta, es probable que su aplicación se expanda, lo que permitirá un análisis más riguroso y matizado de los datos en múltiples disciplinas. Los remuestreos bootstrap no solo han transformado la estadística moderna, sino que encima prometen seguir siendo un pilar fundamental en la investigación futura y en la práctica estadística.

2.3 Medir los índices de fiabilidad, el alfa de Cronbach y los índices de Guttman

La fiabilidad es un concepto fundamental en el ámbito de la investigación y la medición, ya que se refiere a la consistencia y estabilidad de los instrumentos de evaluación utilizados. En términos sencillos, una medición es considerada fiable si produce resultados similares bajo condiciones similares. Esto resulta crucial en investigaciones donde las decisiones y conclusiones dependen de los datos obtenidos.

La fiabilidad se define como la capacidad de un instrumento para medir de manera consistente lo que se pretende medir. Esto implica que, si se repite la medición en las mismas condiciones, los resultados deben ser coherentes y reproducibles (Prieto y Delgado, 2010). La fiabilidad no garantiza que un instrumento mida lo que se propone medir (eso corresponde a la validez), pero sí asegura que los resultados sean consistentes en el tiempo y en diferentes situaciones.

La fiabilidad es un aspecto crítico en la investigación, ya que influye directamente en la validación de los resultados. Un instrumento de medición que carece de fiabilidad consigue conducir a conclusiones erróneas y, por ende, a decisiones inapropiadas basadas en esos datos. En el ámbito académico, la fiabilidad de los instrumentos de medición es esencial para la credibilidad de los estudios y la confianza en las teorías que se desarrollan a partir de ellos. Por añadidura, la fiabilidad es un requisito previo para la validez; es decir, para que un instrumento sea considerado válido, primero debe ser fiable. Existen varios tipos de fiabilidad que los investigadores deben considerar al evaluar sus instrumentos de medición. Entre los más comunes se encuentran:

i. *Fiabilidad test-retest:* Se refiere a la consistencia de los resultados de un mismo instrumento cuando se aplica en dos momentos diferentes a la misma población. Un alto grado de correlación entre las dos mediciones indica una buena fiabilidad test-retest.

ii. *Fiabilidad interjueces:* Evalúa el grado de acuerdo entre diferentes evaluadores o jueces que utilizan el mismo instrumento. Esto es particularmente relevante en estudios cualitativos o en investigaciones donde la interpretación subjetiva consigue influir en los resultados.

iii. *Fiabilidad interna:* Mide la consistencia de los ítems dentro de un mismo instrumento. Se examina cómo se relacionan entre sí las diferentes partes de una prueba o cuestionario. Tanto el alfa de Cronbach como los índices de Guttman son herramientas utilizadas para evaluar esta forma de fiabilidad.

Entender estos tipos de fiabilidad posibilita a los investigadores seleccionar y desarrollar instrumentos de medición que sean más robustos y confiables, lo que a su vez fortalece la calidad de la investigación en su conjunto. El alfa de Cronbach es una de las herramientas más utilizadas para medir la fiabilidad de las escalas y cuestionarios en el ámbito de la investigación social y psicológica. Su popularidad radica en su capacidad para evaluar la consistencia interna de un conjunto de ítems que pretenden medir un mismo constructo.

El alfa de Cronbach se define como un coeficiente que varía entre 0 y 1, donde valores más altos indican una mayor fiabilidad. Se calcula a partir de la varianza de los ítems individuales y la varianza total del conjunto de ítems. La fórmula básica para calcular el alfa de Cronbach (\(\alpha\)) es:

$$
\alpha = \frac{N}{N - 1} \left(1 - \frac{\sum_{i=1}^{N} \sigma^2_{X_i}}{\sigma^2_{X_{total}}}\right)
$$

donde N es el número de ítems, $\sigma^2_{X_i}$ es la varianza de cada ítem y $\sigma^2_{X_{total}}$ es la varianza total del conjunto de ítems. Este cálculo posibilita determinar en qué medida los ítems son coherentes entre sí, proporcionando una indicación de la calidad de la escala.

La interpretación de los valores del alfa de Cronbach no es siempre directa. En general, valores de α superiores a 0.70 se consideran aceptables para investigaciones sociales, si bien algunos expertos sugieren que un valor de 0.80 o más es preferible para asegurar una buena consistencia interna. Empero, un alfa demasiado alto (por encima de 0.95) consigue indicar que los ítems son redundantes y miden esencialmente la misma dimensión. Por lo tanto, es fundamental no solo considerar el valor numérico, sino igualmente el contexto del estudio y la naturaleza de los ítems incluidos.

A pesar de su amplia aceptación, el alfa de Cronbach presenta varias limitaciones, asume que los ítems tienen una estructura unidimensional, lo que significa que miden un solo constructo. Si la escala incluye ítems que abarcan múltiples dimensiones, el alfa consigue ser engañoso. Entonces, el alfa es sensible al número de ítems en la escala; un mayor número de ítems generalmente aumenta el valor del alfa, lo que no necesariamente implica una mejor calidad del instrumento. Por último, el alfa de Cronbach no evalúa la validez del instrumento, por lo que es crucial complementarlo con otras medidas para obtener una visión más integral de la fiabilidad y validez de los instrumentos de medición.

En suma, el alfa de Cronbach es una herramienta valiosa para evaluar la fiabilidad de escalas y cuestionarios. Sin embargo, su uso debe ser acompañado de un análisis crítico y complementado con otros métodos de evaluación para garantizar la robustez de los resultados de la investigación. Los índices de Guttman, también conocidos como escalas de Guttman, son una técnica utilizada para medir la fiabilidad y la unidimensionalidad de un conjunto de ítems o preguntas en una encuesta o cuestionario. Esta metodología se basa en la idea de

que si un individuo posee una cierta característica, es probable que también posea características que se encuentran en niveles inferiores de una jerarquía. Así, los ítems se organizan en una escala que posibilita evaluar la progresión de la presencia de una característica en los encuestados.

En esencia, una escala de Guttman se construye para que, si un participante responde afirmativamente a un ítem de mayor dificultad o complejidad, al igual debería responder afirmativamente a todos los ítems de menor dificultad en esa misma escala. Esto posibilita establecer una jerarquía clara y facilita la interpretación de los resultados. El cálculo de los índices de Guttman implica la creación de una tabla de respuestas, donde cada fila representa a un encuestado y cada columna corresponde a un ítem de la escala. A partir de esta tabla, se consiguen calcular dos tipos de índices: el coeficiente de Guttman (o coeficiente de reproducibilidad) y el índice de consistencia.

El coeficiente de Guttman se define como la proporción de respuestas que se ajustan a la escala establecida. Un valor cercano a 1 indica que la mayoría de las respuestas siguen la jerarquía de la escala, mientras que un valor más bajo sugiere inconsistencias en las respuestas (Engelhard, 2005). Por otro lado, el índice de consistencia posibilita evaluar la homogeneidad de los ítems: si todos los ítems son representativos de la misma dimensión, el índice obtendrá un valor alto.

El uso de los índices de Guttman es especialmente útil en investigaciones donde se busca validar escalas de medición, como en estudios de actitudes, aptitudes o comportamientos. Su aplicación posibilita no solo medir la fiabilidad de las escalas, sino del mismo modo identificar ítems que podrían estar afectando la consistencia general de la medición. Aunque tanto el alfa de Cronbach como los índices de Guttman son herramientas valiosas para evaluar la fiabilidad de las escalas, existen diferencias clave en su planteamiento y aplicación. Mientras que el alfa de Cronbach se basa en la correlación entre ítems y asume que todos los ítems miden la misma construcción subyacente, los índices de Guttman se ajustan en la unidimensionalidad y la jerarquía de respuestas.

El alfa de Cronbach es más adecuado para escalas que no siguen un orden específico, mientras que los índices de Guttman son ideales para escalas que se organizan jerárquicamente. Esta diferencia hace que cada método tenga su lugar dependiendo del tipo de datos y las preguntas de investigación planteadas. A fin

de cuentas, la elección entre el alfa de Cronbach y los índices de Guttman dependerá del contexto de la investigación y de las características de la escala que se está evaluando. La medición de la fiabilidad es un aspecto fundamental en la investigación, ya que garantiza que los instrumentos utilizados para recolectar datos sean consistentes y precisos. La fiabilidad no solo posibilita validar los resultados obtenidos, sino que al igual contribuye a la credibilidad y replicabilidad de los estudios.

El alfa de Cronbach se ha destacado como una de las medidas más comunes para evaluar la fiabilidad interna de un conjunto de ítems, especialmente en cuestionarios y escalas. Su facilidad de cálculo y la interpretación de sus valores proporcionan a los investigadores un recurso accesible para asegurar que sus instrumentos de medición reflejen de manera adecuada el constructo que se desea evaluar (Oviedo y Campo, 2005). Ahora bien, es crucial reconocer sus limitaciones, ya que un alto alfa no siempre garantiza la validez de los ítems ni su capacidad para capturar la complejidad del fenómeno estudiado.

Por otro lado, los índices de Guttman presentan una perspectiva alternativa para evaluar la fiabilidad, especialmente en escalas que se basan en respuestas ordinales. Su planteamiento en la acumulación de ítems posibilita a los investigadores considerar la unidimensionalidad de las medidas, lo que consigue ser particularmente útil en ciertas disciplinas. La comparación entre el alfa de Cronbach y los índices de Guttman revela que, aunque ambos métodos tienen sus propias ventajas y desventajas, su uso conjunto consigue proporcionar una visión más completa de la fiabilidad de los instrumentos.

Así pues, la medición de la fiabilidad es un proceso esencial que debe ser cuidadosamente considerado por los investigadores. La elección entre el alfa de Cronbach y los índices de Guttman, así como el entendimiento de sus implicaciones, consigue influir significativamente en la interpretación de los resultados y en la validez general de la investigación. Por lo tanto, es fundamental que los profesionales del campo se mantengan informados sobre las mejores prácticas y tratamientos actuales en la evaluación de la fiabilidad, asegurando así la calidad y la robustez de sus estudios.

Capítulo III

Análisis de conglomerados. ¿Qué método de agrupación debe elegir?

3.1 Agrupación de K-means, agrupación jerárquica aglomerativa (AHC) y modelos de mezclas gaussianas

La agrupación es una técnica fundamental en el campo de la estadística y el análisis de datos. Su objetivo principal es organizar un conjunto de objetos en grupos o clústeres, de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con aquellos de otros grupos. Esta metodología es especialmente útil en situaciones donde la clasificación de datos no es conocida de antemano y se busca descubrir patrones o estructuras subyacentes en los datos.

La agrupación, todavía conocida como clustering, tiene sus raíces en diversas disciplinas, incluyendo la estadística, la inteligencia artificial y el aprendizaje automático. En términos sencillos, se trata de agrupar datos que presentan características similares. Los algoritmos de agrupación analizan las características de los datos y determinan la mejor manera de organizarlos en clústeres. Estos grupos consiguen ser utilizados para simplificar la representación de los datos, facilitar su análisis o incluso para la identificación de patrones que podrían no ser evidentes de otra manera.

La agrupación juega un papel crucial en el análisis de datos, ya que posibilita a los analistas y científicos de datos explorar grandes volúmenes de información de manera eficiente. Al identificar clústeres en los datos, se consiguen descubrir relaciones ocultas, segmentar mercados, mejorar la personalización de servicios y optimizar procesos en diversas industrias, desde la salud hasta el marketing (Font, 2019). Así, la agrupación consigue servir como un paso preliminar en otros análisis, como la clasificación supervisada, donde se utilizan clústeres para entrenar modelos.

Existen varios métodos y algoritmos para llevar a cabo la agrupación, cada uno con sus propias ventajas y desventajas. Entre los métodos más populares se

encuentran la agrupación de K-means, la agrupación jerárquica aglomerativa (AHC) y los modelos de mezclas gaussianas. Cada una de estas ópticas aborda el problema de la agrupación desde una perspectiva diferente, lo que posibilita a los analistas elegir el método más adecuado según la naturaleza de los datos y los objetivos del análisis. En las secciones siguientes, exploraremos estos métodos en detalle, analizando sus principios, aplicaciones y comparaciones.

El algoritmo K-means es uno de los métodos de agrupación más populares y utilizados en el análisis de datos. Su objetivo principal es dividir un conjunto de datos en un número predefinido de grupos, o "clústeres", que se caracterizan por tener una alta similitud interna y una baja similitud externa con otros grupos. El proceso de K-means se consigue resumir en los siguientes pasos:

i. *Selección de K:* Se elige el número de clústeres (K) que se desea identificar en los datos.

ii. *Inicialización:* Se seleccionan aleatoriamente K puntos de datos como los centroides iniciales de cada clúster.

iii. *Asignación:* Cada punto de datos se asigna al clúster cuyo centroide está más cercano, generalmente utilizando la distancia euclidiana como medida de similitud.

iv. *Actualización:* Una vez que todos los puntos han sido asignados, se recalculan los centroides de los clústeres como la media de los puntos que pertenecen a cada uno.

5. *Iteración:* Los pasos de asignación y actualización se repiten hasta que los centroides no cambian significativamente o se alcanza un número máximo de iteraciones.

El algoritmo K-means ha encontrado aplicación en una variedad de campos y contextos, gracias a su capacidad para identificar patrones en grandes volúmenes de datos. Algunas de las áreas donde se utiliza K-means incluyen:

- *Segmentación de clientes:* En marketing, las empresas utilizan K-means para agrupar a los clientes en función de sus comportamientos de compra, lo que les posibilita personalizar las estrategias de marketing y mejorar la satisfacción del cliente.

- *Análisis de imágenes:* En procesamiento de imágenes, K-means se emplea para la segmentación de imágenes, permitiendo la identificación de diferentes regiones o elementos dentro de una imagen.

- *Agrupación de documentos:* En el ámbito del procesamiento de lenguaje natural, K-means se utiliza para agrupar documentos similares, facilitando la organización y recuperación de información.

En síntesis, el algoritmo K-means es una técnica poderosa de agrupación que, a pesar de sus limitaciones, ha demostrado ser eficaz en una amplia gama de aplicaciones prácticas. Su simplicidad y velocidad lo convierten en una opción preferida para muchos analistas de datos. La agrupación jerárquica aglomerativa (AHC, por sus siglas en inglés) es una técnica de análisis de datos que posibilita organizar un conjunto de objetos en una jerarquía de grupos o clústeres. A diferencia de métodos de agrupación como K-means que requieren especificar el número de clústeres de antemano, la AHC construye una jerarquía que consigue ser visualizada mediante un dendrograma, permitiendo una exploración más flexible de la estructura de datos.

El proceso de AHC comienza considerando cada objeto como un clúster individual, y existen diferentes métodos para calcular esta distancia, como la distancia euclidiana, la distancia Manhattan o la correlación. La AHC consigue ser implementada utilizando dos planteamientos principales:

i. *Enlace sencillo (single linkage):* Se une el par de clústeres que tienen la menor distancia entre sus puntos más cercanos.

ii. *Enlace completo (complete linkage):* Se une el par de clústeres que tienen la menor distancia entre sus puntos más lejanos.

Este proceso se repite hasta que todos los objetos se combinan en un solo clúster o se alcanza un número predefinido de clústeres, lo que posibilita una visualización clara de la estructura jerárquica de los datos. Una de las diferencias más significativas entre AHC y K-means es la forma en que ambos métodos definen los clústeres. K-means asigna los puntos a clústeres basándose en la proximidad a centroides, que son recalculados en cada iteración. En contraste, la AHC no depende de centroides, sino que se basa en la distancia entre los objetos y combina clústeres en función de la similitud.

Otra diferencia clave es la determinación del número de clústeres. En K-means, el número de clústeres debe ser especificado antes de la ejecución del algoritmo, mientras que la AHC posibilita una exploración más flexible del número de clústeres a través de la visualización del dendrograma, donde se consigue elegir el nivel de agrupación más adecuado observando la altura de los enlaces.

La agrupación jerárquica aglomerativa se utiliza en diversos campos debido a su capacidad para revelar estructuras ocultas en los datos. En biología, se emplea para clasificar especies basándose en similitudes genéticas. En la segmentación de mercado, las empresas consiguen usar AHC para identificar grupos de consumidores con comportamientos similares, lo que les posibilita personalizar sus estrategias de marketing (Yadav y Dhull, 2024). En esa misma línea, en el análisis de texto, AHC consigue ser utilizada para agrupar documentos similares, facilitando la organización y recuperación de información. Este planteamiento es especialmente útil en el procesamiento de lenguaje natural, donde se busca identificar temas o categorías dentro de grandes volúmenes de texto.

Como se ha dicho, la agrupación jerárquica aglomerativa es una herramienta valiosa en el análisis de datos, proporcionando una forma intuitiva de explorar y visualizar la estructura de los datos sin la necesidad de especificar de antemano el número de clústeres. Los modelos de mezclas gaussianas (GMM, por sus siglas en inglés) son una técnica estadística utilizada para modelar la distribución de datos a través de la combinación de múltiples distribuciones gaussianas.

Cada una de estas distribuciones representa un "grupo" o "clúster" en el conjunto de datos, y se caracterizan por su media y su covarianza. La idea central es que, en lugar de asumir que todos los datos provienen de una única distribución, se postula que los datos son generados por un número de distribuciones gaussianas, cada una con sus propios parámetros, y que estas combinaciones consiguen ser utilizadas para describir la complejidad de la estructura de los datos.

Los GMM se basan en la teoría de que los datos consiguen ser representados como una mezcla de varias distribuciones gaussianas, donde cada componente de la mezcla tiene su propia probabilidad de ocurrencia. Este

planteamiento posibilita que el modelo capture la variabilidad y la complejidad de los datos de una manera más flexible en comparación con métodos más simples, como el K-means, que asume que todos los clústeres son esféricos y de igual tamaño.

Entre las principales diferencias entre los modelos de mezclas gaussianas y los métodos de agrupación como K-means y la agrupación jerárquica aglomerativa (AHC) radica en la forma en que se asignan las instancias a los clústeres. En K-means, cada punto de datos se asigna de manera determinista al clúster más cercano, lo que implica que las fronteras entre los clústeres son rígidas. En contraste, los GMM posibilitan que un punto de datos pertenezca a múltiples clústeres con diferentes grados de pertenencia, lo que resulta en una representación más suave y realista de la estructura de los datos.

En suma, mientras que AHC crea clústeres jerárquicos y se basa en distancias entre puntos, los GMM utilizan un planteamiento probabilístico que considera las características de cada clúster y su distribución. Esto significa que los GMM consiguen adaptarse mejor a la forma y tamaño de los clústeres, lo cual es particularmente útil en situaciones donde los datos no son esféricos o tienen una alta dimensionalidad.

Los modelos de mezclas gaussianas han demostrado ser extremadamente útiles en una variedad de aplicaciones prácticas. En el campo de la visión por computadora, se utilizan para la segmentación de imágenes, donde diferentes áreas de una imagen consiguen ser modeladas como distintas mezclas gaussianas (Reynolds, 2009). Esto posibilita distinguir entre diferentes objetos y texturas dentro de una imagen de manera efectiva.

En el ámbito de la biología, los GMM se han empleado para clasificar diferentes especies o grupos de organismos basándose en características morfológicas o genéticas, permitiendo a los investigadores identificar patrones en los datos que podrían no ser evidentes a través de métodos más simples. Asimismo, en el análisis de mercados, estas técnicas se utilizan para segmentar consumidores en grupos con comportamientos similares, facilitando la personalización de estrategias de marketing.

Dicho de otro modo, los modelos de mezclas gaussianas dedican un planteamiento robusto y flexible para la agrupación y el modelado de datos complejos, destacándose en situaciones donde las suposiciones de otros métodos

de agrupación consiguen no ser adecuadas. Su capacidad para manejar la incertidumbre y la variabilidad en los datos los convierte en una herramienta valiosa en el análisis estadístico moderno. El análisis clúster presenta ópticas únicas para el agrupamiento de datos, permitiendo a los analistas identificar patrones y estructuras dentro de conjuntos de datos complejos. K-means se destaca por su simplicidad y velocidad, siendo ideal para grandes volúmenes de datos, mientras que AHC proporciona una visualización jerárquica que facilita la comprensión de las relaciones entre grupos. Por otro lado, los modelos de mezclas gaussianas ofrecen un marco probabilístico que posibilita una mayor flexibilidad en la asignación de datos a grupos, capturando mejor la incertidumbre inherente en el análisis.

La elección de la técnica de agrupación adecuada depende de diversos factores, incluyendo la naturaleza de los datos, los objetivos del análisis y la interpretación deseada de los resultados. Si se requiere una agrupación rápida y eficiente en grandes conjuntos de datos, K-means consigue ser la opción más viable. En contraste, si el objetivo es comprender la estructura jerárquica de los datos, AHC podría ser más apropiado. Los modelos de mezclas gaussianas son particularmente útiles en contextos donde se necesita modelar la variabilidad y la superposición de grupos. Es fundamental que los analistas consideren estas características al seleccionar la técnica más adecuada para sus necesidades específicas, así como la posibilidad de combinar métodos para obtener resultados más robustos.

A medida que el campo del análisis de datos continúa evolucionando, también lo hacen las técnicas de agrupación. Investigaciones futuras podrían enfocarse en el desarrollo de algoritmos más eficientes que integren las ventajas de las técnicas existentes, así como en la mejora de la interpretabilidad de los resultados. Por añadidura, la creciente disponibilidad de datos masivos y complejos plantea nuevos desafíos y oportunidades para la agrupación, impulsando la necesidad de métodos que sean tanto escalables como adaptativos.

La incorporación de técnicas de aprendizaje automático e inteligencia artificial en el proceso de agrupación al igual representa un área prometedora, permitiendo la creación de modelos que no solo agrupen datos, sino que también aprendan y se adapten a medida que surgen nuevos patrones en los datos. A fin

de cuentas, el futuro de la agrupación en el análisis de datos es emocionante y está lleno de posibilidades, lo que promete acumular la comprensión y el uso de datos en diversas disciplinas.

3.2 Agrupación univariante y modelos de agrupación de clases latentes

La agrupación univariante es una técnica estadística utilizada para clasificar datos en grupos homogéneos basándose en una única variable. Su simplicidad y facilidad de interpretación la convierten en una herramienta valiosa en diversas disciplinas, desde la psicología hasta la biología. A través de esta perspectiva, los investigadores consiguen identificar patrones y tendencias que de otro modo podrían pasar desapercibidos. Pese a, la naturaleza univariante limita la riqueza de la información que se consigue extraer, ya que ignora las complejidades que consiguen surgir de múltiples variables interrelacionadas.

En este contexto, los modelos de agrupación de clases latentes emergen como una solución poderosa. Estos modelos posibilitan la identificación de subgrupos en una población a partir de datos observables, considerando múltiples variables simultáneamente. Esto enriquece el análisis y proporciona una comprensión más profunda de la estructura subyacente de los datos. Al modelar la heterogeneidad de la población, los investigadores consiguen descubrir clases latentes que representan patrones de comportamiento o características no observadas directamente.

La importancia de los modelos de agrupación de clases latentes radica en su capacidad para captar la complejidad del comportamiento humano y social. A medida que avanzan las técnicas de análisis de datos y la informática, la integración de estos modelos se convierte en un componente crucial en la investigación moderna. La habilidad para identificar y caracterizar grupos heterogéneos no solo mejora la precisión de las inferencias estadísticas, sino que también tiene implicaciones prácticas en áreas como la segmentación de mercados, el desarrollo de políticas públicas, y la personalización de tratamientos en el ámbito de la salud.

Para Quero e Inciarte (2012), la agrupación univariante es una técnica estadística que se focaliza en la clasificación de datos en grupos o clústeres basándose en una sola variable. A diferencia de la agrupación multivariante, que

considera múltiples dimensiones o variables simultáneamente, la agrupación univariante simplifica el análisis al enfocarse en un solo aspecto del conjunto de datos. Esta metodología es especialmente útil cuando se busca comprender patrones o tendencias dentro de un único atributo, facilitando la identificación de grupos homogéneos. La agrupación univariante se define como el proceso de segmentar un conjunto de datos en grupos que comparten características similares, basándose exclusivamente en la variabilidad de una única variable.

Esta técnica posibilita a los investigadores y analistas realizar categorizaciones que consiguen ser cruciales para el entendimiento de fenómenos específicos. En un estudio sobre las alturas de una población, la agrupación univariante podría emplearse para identificar grupos de personas con alturas similares, sin considerar otros factores como el peso o la edad. Las características más destacadas de la agrupación univariante incluyen su simplicidad y la facilidad de interpretación de los resultados.

Al concentrarse en una sola variable, los resultados son más directos y comprensibles, lo que posibilita a los investigadores presentar sus hallazgos de manera clara. Esta técnica es ampliamente utilizada en diversas áreas, como la biología para clasificar especies según características morfológicas, en la psicología para identificar perfiles de personalidad basados en pruebas univariadas, y en el marketing para segmentar clientes según una variable clave como la edad o el ingreso. Algunas aplicaciones concretas incluyen el uso de histogramas y diagramas de caja para visualizar la distribución de datos y la identificación de outliers, así como la aplicación de métodos como el k-means o el análisis de conglomerados jerárquico para agrupar datos basados en una única medida.

A pesar de sus ventajas, la agrupación univariante tiene limitaciones significativas. La más notable es que, al considerar solo una variable, consigue perderse información crucial que podría ser relevante para la clasificación. En la segmentación de mercado, una agrupación basada únicamente en la edad consigue no capturar la complejidad del comportamiento del consumidor, que consigue estar influenciado por múltiples factores, como los intereses o el contexto socioeconómico.

Al respecto, la agrupación univariante consigue ser susceptible a la variabilidad aleatoria en los datos y consigue resultar en agrupaciones que no

reflejan patrones subyacentes más complejos. Esto consigue llevar a conclusiones erróneas si no se complementa con un análisis más exhaustivo que considere múltiples variables. En suma, la agrupación univariante es una herramienta valiosa en el análisis de datos que posibilita identificar grupos homogéneos basándose en una sola variable. No obstante, es fundamental ser consciente de sus limitaciones y considerar su uso como un primer paso hacia un análisis más profundo que incluya múltiples dimensiones.

Los modelos de clases latentes (MCL) son herramientas estadísticas que posibilitan identificar subgrupos o clases dentro de un conjunto de datos, basándose en patrones de respuesta observados. A diferencia de los métodos tradicionales de agrupación, que consiguen depender de variables observadas explícitamente, los MCL se orientan en las estructuras subyacentes que consiguen no ser evidentes a simple vista (Sinha et al., 2021). Estos modelos son especialmente útiles en situaciones donde los datos consiguen ser heterogéneos y se desea descubrir la existencia de grupos que comparten características comunes, a pesar de que no se disponga de información directa sobre las clasificaciones de los casos.

Existen diversos tipos de modelos de clases latentes, cada uno adaptado a diferentes tipos de datos y objetivos de investigación. Algunos de los más comunes incluyen:

i. *Modelos de Clases Latentes Finitos:* Estos modelos asumen que los datos se consiguen dividir en un número finito de clases latentes. Cada clase se caracteriza por un patrón de respuesta distinto, y los individuos se asignan a estas clases en función de su probabilidad de pertenencia.

ii. *Modelos de Clases Latentes Mixtos:* A diferencia de los modelos finitos, estos posibilitan la existencia de subgrupos dentro de las clases latentes, lo que facilita la modelización de datos más complejos donde las características de los grupos consiguen solaparse.

iii. *Modelos de Clases Latentes Estructurales:* Estos modelos combinan la identificación de clases latentes con variables latentes que consiguen influir en las respuestas observadas, permitiendo una comprensión más profunda de las relaciones subyacentes.

iv. *Modelos de Clases Latentes de Rasch:* Utilizados principalmente en psicometría, estos modelos posibilitan la evaluación de rasgos latentes a partir de respuestas a ítems en pruebas o cuestionarios, asumiendo que las respuestas son función de la habilidad del individuo y la dificultad de los ítems.

Los modelos de clases latentes presentan varias ventajas significativas en el ámbito de la agrupación:

i. *Flexibilidad:* Consiguen adaptarse a diferentes tipos de datos (categóricos, continuos, ordinales) y a diversas estructuras de datos, lo que los hace útiles en múltiples disciplinas.

ii. *Identificación de Estructuras Ocultas:* Posibilitan descubrir patrones de agrupación que no son evidentes a través de análisis descriptivos o métodos de agrupación más simples.

iii. *Interpretación de Resultados:* Los resultados obtenidos a partir de MCL son frecuentemente más fáciles de interpretar, ya que proporcionan información sobre las características de las clases y la probabilidad de pertenencia de los individuos a cada una de ellas.

iv. *Manejo de la Incertidumbre:* Los MCL posibilitan modelar la incertidumbre en la clasificación, lo que es particularmente útil en contextos donde las fronteras entre grupos no son claras.

v. *Aplicabilidad a Datos Complejos:* Son especialmente valiosos en el análisis de datos complejos, como encuestas con múltiples ítems o datos longitudinales, donde las relaciones entre variables consiguen ser intrincadas.

En otras palabras, los modelos de clases latentes ofrecen una metodología robusta y versátil para la agrupación de datos, permitiendo a los investigadores explorar y entender mejor las dinámicas subyacentes en sus áreas de estudio. La agrupación univariante y los modelos de agrupación de clases latentes han encontrado un amplio espectro de aplicaciones en diversas disciplinas de investigación. Estas técnicas posibilitan a los investigadores descomponer datos complejos en segmentos más manejables, facilitando así la identificación de patrones y relaciones subyacentes.

En el ámbito de la psicología, la agrupación univariante se utiliza para clasificar a los individuos en función de características específicas, como rasgos

de personalidad, niveles de ansiedad o patrones de comportamiento. Un estudio podría utilizar la agrupación univariante para identificar grupos de pacientes con trastornos de ansiedad que presentan síntomas similares, lo que podría facilitar el desarrollo de tratamientos más personalizados. Por otro lado, los modelos de clases latentes son especialmente útiles en investigaciones sociales para identificar grupos ocultos en la población. Un ejemplo sería el análisis de datos de encuestas para segmentar a los encuestados según sus actitudes hacia temas sociales como la inmigración o el cambio climático. Estas clases latentes posibilitan a los investigadores comprender mejor la heterogeneidad de las opiniones dentro de la población, lo que consigue influir en la formulación de políticas públicas.

En el campo del marketing, las técnicas de agrupación univariante y los modelos de clases latentes son herramientas valiosas para segmentar mercados y entender el comportamiento del consumidor. La agrupación univariante posibilita a las empresas identificar grupos de consumidores con características similares, como preferencias de compra o hábitos de consumo, lo que ayuda a personalizar las estrategias de marketing.

Los modelos de clases latentes, en particular, ofrecen la capacidad de descubrir segmentos de consumidores que no son evidentes a simple vista. Se consiguen aplicar para identificar diferentes perfiles de consumidores en función de sus respuestas a encuestas sobre productos, revelando así segmentos de mercado que consiguen ser dirigidos de manera más efectiva con campañas publicitarias específicas.

En biología y estudios ecológicos, la agrupación univariante y los modelos de clases latentes se utilizan para clasificar especies o poblaciones en función de características biológicas o ambientales. En estudios de biodiversidad, la agrupación univariante consigue ayudar a identificar grupos de especies que comparten características ecológicas similares, facilitando así la comprensión de la dinámica de los ecosistemas.

Los modelos de clases latentes en esa misma línea son aplicados en la ecología para identificar grupos de hábitats o comunidades biológicas que presentan patrones similares en su composición. Esto es crucial para la conservación de la biodiversidad, ya que posibilita a los ecólogos enfocar sus esfuerzos en áreas que albergan comunidades vulnerables o en peligro. En

síntesis, las aplicaciones de la agrupación univariante y los modelos de clases latentes son diversas y abarcan múltiples disciplinas. Desde la psicología hasta el marketing y la biología, estas técnicas proporcionan herramientas poderosas para el análisis y la interpretación de datos, permitiendo a los investigadores desentrañar la complejidad de comportamientos y patrones en sus respectivos campos

La agrupación univariante y los modelos de agrupación de clases latentes representan dos planteamientos fundamentales en el análisis de datos que han evolucionado de manera significativa en las últimas décadas. Su relevancia se manifiesta en su capacidad para desglosar y comprender la complejidad de los datos, permitiendo a los investigadores identificar patrones subyacentes y estructuras latentes que de otro modo podrían pasar desapercibidos (Batanero, 2001).

A medida que la cantidad de datos disponibles continúa creciendo exponencialmente, la necesidad de técnicas robustas y eficientes para la agrupación se vuelve cada vez más apremiante. La agrupación univariante, aunque limitada en su capacidad para capturar la complejidad multidimensional de las variables, sigue siendo una herramienta valiosa en contextos donde se busca una interpretación simple y directa. Su aplicación en diversas disciplinas, desde la psicología hasta la biología, subraya su utilidad en la extracción de información relevante de conjuntos de datos específicos.

Por otro lado, los modelos de agrupación de clases latentes han cambiado el paradigma del análisis de datos al permitir una mayor flexibilidad y precisión en la identificación de subgrupos dentro de la población. Estos modelos no solo ofrecen ventajas a nivel metodológico, sino que siempre brindan una comprensión más profunda del comportamiento y las características de los individuos dentro de esos subgrupos. Su capacidad para manejar datos faltantes y su aplicabilidad en contextos donde las relaciones entre variables son complejas los convierten en una herramienta esencial para los investigadores.

Mirando hacia el futuro, es probable que la integración de técnicas de agrupación univariante y modelos de clases latentes continúe desarrollándose, especialmente con el avance de tecnologías de análisis de datos y aprendizaje automático. La combinación de planteamientos tradicionales y nuevas metodologías permitirá a los científicos y profesionales abordar preguntas de

investigación cada vez más complejas y multidimensionales. Tanto la agrupación univariante como los modelos de agrupación de clases latentes tienen un papel crucial en la investigación actual y futura. Su evolución y adaptación a nuevas realidades de datos son esenciales para seguir proporcionando lo que sin duda influirá en la forma en que entendemos y analizamos el comportamiento humano, las tendencias del mercado y los fenómenos ecológicos en un mundo en constante cambio.

3.3 Análisis de correspondencias múltiples (MCA)

El análisis de correspondencias múltiples (MCA) es una poderosa técnica estadística que se utiliza para explorar las relaciones dentro de los datos categóricos. Amplía el concepto de análisis de correspondencias, que se orienta en examinar dos variables categóricas, lo que posibilita a los investigadores visualizar e interpretar conjuntos de datos complejos que incluyen múltiples variables categóricas (Sourial et al., 2010). Las raíces del MCA se encuentran en el análisis multivariante, donde sirve como una herramienta valiosa para descubrir patrones y asociaciones que consiguen no ser inmediatamente evidentes a través de los métodos tradicionales de análisis de datos.

En esencia, el análisis de correspondencias múltiples está diseñado para analizar y representar datos categóricos en un espacio de dimensiones inferiores, lo que facilita la interpretación y la visualización. Al transformar los datos en un conjunto de dimensiones, el MCA ayuda a los investigadores a identificar las estructuras subyacentes, las asociaciones y las similitudes entre las categorías. La representación gráfica resultante, normalmente en forma de diagrama de dispersión, posibilita a los usuarios ver cómo se relacionan las diferentes categorías entre sí y observar la distribución de las observaciones en estas dimensiones.

El ACM desempeña un papel crucial en el análisis de datos, especialmente en campos donde prevalecen los datos categóricos. Su importancia radica en su capacidad para simplificar conjuntos de datos complejos, lo que facilita a los analistas la obtención de información y la toma de decisiones informadas. Al revelar patrones y relaciones, MCA apoya la generación de hipótesis, la exploración de datos y la validación de teorías. Además, sus resultados visuales mejoran la comunicación de los resultados, lo que posibilita a las partes interesadas comprender relaciones complejas de forma intuitiva.

La versatilidad de MCA ha llevado a su adopción en varios campos y aplicaciones. En la investigación de mercado, a menudo se emplea para comprender las preferencias y comportamientos de los consumidores mediante el análisis de datos de encuestas. En las ciencias sociales, los investigadores utilizan el ACM para explorar las relaciones entre las variables demográficas y las actitudes sociales. En especial, los estudios ambientales aprovechan el MCA para examinar los datos ecológicos, lo que ayuda a identificar los factores que influyen en la biodiversidad y la distribución del hábitat. Como resultado, MCA sirve como una herramienta analítica vital para los investigadores que buscan obtener información significativa de datos categóricos en diversos dominios.

Para apreciar plenamente las capacidades y aplicaciones del Análisis de Correspondencias Múltiples (ACM), es imprescindible profundizar en sus fundamentos teóricos. Estos abarcan los principios estadísticos subyacentes, las comparaciones con metodologías relacionadas y la comprensión de sus supuestos y limitaciones. MCA amplía el concepto de Análisis de Correspondencias (CA), diseñado específicamente para datos categóricos. Proporciona una forma de analizar y visualizar las relaciones entre múltiples variables categóricas transformándolas en un espacio de menor dimensión, conservando la estructura de datos original tanto como sea posible. El principio fundamental del MCA consiste en crear una tabla de contingencia a partir de datos categóricos, en la que las filas representan individuos u observaciones, y las columnas representan las categorías de las variables.

A través de la descomposición de valores singulares (SVD), el MCA identifica las estructuras subyacentes dentro de los datos, lo que posibilita a los investigadores examinar patrones y asociaciones entre categorías. Las dimensiones producidas por MCA representan variables latentes que capturan la varianza presente en el conjunto de datos, proporcionando información sobre las relaciones entre diferentes variables categóricas.

Si bien tanto el MCA como el análisis de componentes principales (PCA) son técnicas de reducción de dimensiones que se utilizan para descubrir estructuras subyacentes en los datos, se adaptan a diferentes tipos de datos y sirven para fines distintos. El PCA es adecuado principalmente para datos numéricos continuos, centrándose en maximizar la varianza a través de combinaciones lineales de variables. Por el contrario, el MCA se adapta a datos

categóricos, haciendo hincapié en las asociaciones y relaciones entre categorías en lugar de maximizar la varianza.

La interpretación geométrica del ACP implica proyectar puntos de datos en un espacio de menor dimensión donde se maximiza la varianza. Por el contrario, en MCA, la atención se canaliza en las relaciones entre categorías, visualizadas a través de un mapa perceptual que ilustra la proximidad de las categorías en función de sus asociaciones. Esta diferencia fundamental pone de manifiesto la capacidad de MCA para revelar información que consigue no ser evidente a través de la PCA tradicional, especialmente cuando se trata de conjuntos de datos categóricos complejos.

A pesar de sus fortalezas, el MCA opera bajo un conjunto de supuestos que los investigadores deben considerar al aplicar el método. Un supuesto primario es la independencia de las observaciones; el ACM supone que las categorías analizadas son independientes entre sí. De igual manera, la técnica asume que las relaciones entre las categorías son lineales, lo que no siempre es cierto en los datos del mundo real.

Otra limitación de MCA es su sensibilidad al tamaño del conjunto de datos. Los conjuntos de datos grandes consiguen producir resultados más significativos, mientras que los conjuntos de datos más pequeños consiguen dar lugar a interpretaciones menos fiables. Por añadidura, el MCA consigue llegar a ser computacionalmente intensivo, particularmente con datos de alta dimensión, lo que requiere el uso de software especializado y recursos computacionales avanzados.

Comprender los fundamentos teóricos del ACM es crucial para los investigadores que desean implementar esta técnica de manera efectiva en sus análisis. Al comprender los antecedentes estadísticos, distinguirlos de los métodos relacionados como el PCA y reconocer sus suposiciones y limitaciones, los analistas consiguen aprovechar el MCA para descubrir información valiosa dentro de conjuntos de datos categóricos (Lamfre et al., 2023). El análisis de correspondencias múltiples (MCA) es una herramienta poderosa para analizar datos categóricos, pero su efectividad depende de una implementación cuidadosa.

El primer paso para implementar el MCA es asegurarse de que los datos se preparen adecuadamente. El MCA está diseñado específicamente para

variables categóricas, por lo que el conjunto de datos debe constar de tipos de datos cualitativos. Estas son algunas consideraciones clave para la preparación de datos:

i. *Recopilación de datos:* Recopilar datos de fuentes confiables, asegurándose de que las variables de interés sean categóricas. Esto podría incluir respuestas a encuestas, información demográfica o cualquier dato que se pueda clasificar en distintas categorías.

ii. *Manejo de datos faltantes:* Los valores faltantes consiguen afectar significativamente los resultados de MCA. Las opciones incluyen métodos de imputación, en los que los valores faltantes se estiman sobre la base de otras observaciones, o excluyen del análisis a individuos con datos incompletos.

iii. *Recodificación de variables:* Es posible que sea necesario recodificar las variables para asegurarse de que estén en un formato adecuado para MCA. Se trata de transformar las variables en factores o variables ficticias, especialmente si se presentan como datos numéricos.

iv. *Estandarización:* Si bien el MCA no requiere el mismo nivel de estandarización que técnicas como el PCA, es crucial garantizar que las categorías se definan de manera consistente. Si una variable tiene respuestas como "sí" y "no", deben codificarse de manera uniforme en todo el conjunto de datos.

v. *Creación de una tabla de contingencia:* Antes de realizar el MCA, consigue ser útil crear una tabla de contingencia que resuma las relaciones entre las variables categóricas. Esta tabla sirve como base para el análisis de MCA.

Varios paquetes de software y herramientas consiguen facilitar la ejecución de MCA. Cada uno tiene sus puntos fuertes, y la elección de la herramienta consigue depender de las preferencias del usuario, la complejidad del análisis o el conjunto de datos específico. Algunas opciones populares incluyen:

i. *R:* El lenguaje de programación R presenta varios paquetes para realizar MCA, como 'FactoMineR' y 'ca'. Estas herramientas proporcionan funciones completas para realizar MCA y visualizar los resultados, lo que hace que R sea una opción popular entre los analistas de datos.

ii. *Python:* Python encima tiene bibliotecas como 'Prince' y 'scikit-learn' que consiguen realizar MCA. Estas bibliotecas son particularmente atractivas para los usuarios familiarizados con el ecosistema de Python, ya que se integran bien con otras herramientas de manipulación y análisis de datos.

iii. *SPSS:* SPSS presenta una interfaz fácil de usar para realizar MCA, lo que lo hace accesible para aquellos que consiguen no tener experiencia en programación. El software proporciona funcionalidades integradas para el preprocesamiento, análisis y visualización de datos.

iv. *SAS:* SAS incluye procedimientos para MCA, lo que posibilita a los usuarios realizar análisis en profundidad de datos categóricos. Sus sólidas capacidades lo hacen adecuado para conjuntos de datos más grandes y análisis complejos.

v. *Excel:* Si bien no está diseñado específicamente para MCA, Excel se consigue usar para el análisis básico de MCA a través de complementos o cálculos manuales. Empero, este planteamiento consigue ser limitado en términos de escalabilidad y funcionalidades avanzadas. Una vez que se lleva a cabo el MCA, la interpretación de los resultados es esencial para obtener información significativa. Los aspectos clave a tener en cuenta incluyen:

i. *Valores propios y dimensiones:* La salida suele incluir valores propios, que indican la varianza capturada por cada dimensión. Los valores propios más altos sugieren un mayor poder explicativo, y los analistas deben centrarse en las dimensiones que representan la mayor varianza.

ii. *Mapas de factores:* Los resultados de MCA a menudo se visualizan utilizando mapas de factores, que trazan las categorías de variables en un espacio bidimensional. Esta visualización ayuda a identificar relaciones y grupos entre las categorías, revelando patrones y asociaciones que consiguen no ser evidentes de inmediato.

iii. *Contribución de Variables:* Se consigue analizar la contribución de cada variable a las dimensiones. Esta información ayuda a comprender qué variables son las más influyentes en la diferenciación de las categorías.

iv. *Calidad de la representación:* Es crucial evaluar la calidad de la representación tanto para las categorías como para los individuos en las parcelas resultantes. Las métricas, como el coseno cuadrado para las categorías y las coordenadas para los

individuos, proporcionan información sobre qué tan bien están representados en las dimensiones reducidas.

v. *Análisis de conglomerados:* Después de obtener los resultados del MCA, los analistas consiguen realizar análisis de conglomerados para agrupar categorías similares en función de su proximidad en el espacio de factores. Este paso consigue mejorar la comprensión de los patrones subyacentes en los datos.

La implementación práctica de MCA implica una preparación cuidadosa de los datos, la selección de las herramientas adecuadas y la interpretación efectiva de los resultados. Al seguir estas pautas, los investigadores y analistas consiguen aprovechar el MCA para descubrir información significativa a partir de datos categóricos en varios campos. (Lamfre et al., 2023) El Análisis de Correspondencias Múltiples (MCA) se ha convertido en una poderosa herramienta para descubrir patrones en datos categóricos en varios campos.

En el ámbito de la investigación de mercado, el MCA se utiliza ampliamente para comprender las preferencias y comportamientos de los consumidores. Un estudio de caso destacado involucró a una empresa minorista líder que realizó un MCA para analizar los comentarios de los clientes recopilados a través de encuestas. El conjunto de datos incluía múltiples variables categóricas, como el grupo de edad, el sexo, la frecuencia de compra y la preferencia de producto. Al aplicar MCA, los investigadores pudieron identificar distintos segmentos de consumidores y sus preferencias asociadas, revelando información sobre el posicionamiento de la marca en el mercado.

Los resultados indicaron que los consumidores más jóvenes preferían los productos ecológicos, mientras que los grupos demográficos de mayor edad mostraban una preferencia por las marcas tradicionales. Esta comprensión matizada permitió a la empresa adaptar sus estrategias de marketing de manera efectiva, dirigiéndose a segmentos específicos con campañas personalizadas. La capacidad de MCA para visualizar relaciones complejas entre variables categóricas lo convirtió en un activo invaluable para elaborar estrategias de ofertas de productos y mejorar la satisfacción del cliente.

El MCA también ha encontrado aplicaciones significativas en las ciencias sociales, donde los investigadores a menudo tratan con datos categóricos multivariados. Un ejemplo ilustrativo proviene de un estudio sociológico que explora la relación entre el nivel educativo, la situación laboral y la movilidad

social. Los investigadores recopilaron datos de una población diversa, que abarcaba varias variables demográficas como el origen étnico, el nivel de ingresos y la ubicación geográfica.

A través de MCA, los investigadores crearon una representación visual completa de los datos, revelando grupos de individuos con perfiles similares. El análisis demostró que los niveles más altos de educación estaban fuertemente asociados con mayores oportunidades de empleo, particularmente entre los grupos minoritarios. Esta visión no solo contribuyó a las discusiones académicas sobre la equidad social, sino que al igual informó a los responsables de la formulación de políticas que buscan diseñar programas educativos efectivos para cerrar la brecha de movilidad.

Los estudios ambientales y la ecología también se han beneficiado de la aplicación de la MCA, particularmente en el análisis de estudios ecológicos y la comprensión de la distribución de las especies. Un estudio de caso se centró en la evaluación de la biodiversidad en una región específica en la que los investigadores recopilaron datos categóricos sobre la presencia de especies, los tipos de hábitat y las condiciones ambientales.

Mediante el empleo de MCA, los investigadores identificaron patrones en la distribución de las especies y sus relaciones con diversos factores ambientales. El análisis reveló que ciertas especies prosperaban en hábitats específicos, mientras que otras eran más adaptables a las condiciones cambiantes. Este conocimiento fue crucial para los esfuerzos de conservación, guiando las estrategias para proteger las especies vulnerables y restaurar el equilibrio ecológico.

Entonces, MCA facilitó la visualización de interacciones complejas entre las especies y sus hábitats, lo que facilitó que las partes interesadas comunicaran los hallazgos y participaran en iniciativas de conservación colaborativas. En teoría, las diversas aplicaciones de MCA en la investigación de mercados, las ciencias sociales y los estudios ambientales subrayan su importancia como una herramienta analítica sólida. Estos estudios de caso ilustran cómo el MCA no solo mejora la comprensión de los datos categóricos, sino que también informa los procesos de toma de decisiones en varios sectores, lo que en última instancia conduce a estrategias y soluciones más efectivas para abordar problemas complejos.

El Análisis de Correspondencias Múltiples (MCA) es una poderosa herramienta estadística que facilita la exploración y visualización de datos categóricos. Al permitir a los investigadores identificar patrones y relaciones entre variables, el MCA proporciona información valiosa que consigue impulsar la toma de decisiones en varios campos. Los principios fundamentales del MCA, arraigados en la teoría estadística, subrayan su relevancia y eficacia en comparación con otras técnicas como el Análisis de Componentes Principales (PCA).

Como se ha comentado, la implementación práctica de MCA requiere una cuidadosa preparación de datos y el uso de herramientas de software adecuadas. La capacidad de interpretar los resultados de MCA a través de visualizaciones efectivas garantiza que los hallazgos sean accesibles y procesables para diversas audiencias, lo que mejora la utilidad de este método analítico. En suma, los estudios de caso explorados demuestran la versatilidad del MCA, mostrando su aplicación en la investigación de mercados, las ciencias sociales y los estudios ambientales. Estos ejemplos ponen de manifiesto cómo el MCA consigue revelar relaciones complejas dentro de los datos, lo que conduce a estrategias y políticas más informadas.

De cara al futuro, los posibles avances en MCA, incluidas las técnicas computacionales mejoradas y la integración con algoritmos de aprendizaje automático, prometen ampliar su aplicabilidad y mejorar su precisión. A medida que los investigadores continúen buscando formas innovadoras de analizar los datos, la importancia del MCA para proporcionar una comprensión matizada de las variables categóricas sin duda aumentará. El Análisis de Correspondencias Múltiples se erige como una herramienta vital en la investigación contemporánea, que posibilita tanto a los académicos como a los profesionales descubrir ideas significativas a partir de datos categóricos. Su desarrollo y aplicación continuos seguirán desempeñando un papel crucial en la configuración de las metodologías de investigación en diversas disciplinas.

3.4 Ejecución de una agrupación jerárquica aglomerativa (AHC) tras un MCA

La agrupación jerárquica aglomerativa (AHC, por sus siglas en inglés) es una técnica fundamental en el campo del análisis de datos y la minería de datos, utilizada para identificar patrones y estructuras en conjuntos de datos complejos.

Esta metodología posibilita organizar un conjunto de objetos o datos en grupos o clústeres jerárquicos, proporcionando una visualización clara de las relaciones y similitudes entre los elementos analizados. A medida que los volúmenes de datos continúan creciendo en diversas disciplinas, desde la biología hasta el marketing, la capacidad de agrupar información de manera efectiva se vuelve cada vez más crucial.

El proceso de agrupación jerárquica aglomerativa comienza con cada objeto identificado como un clúster individual. A partir de ahí, los clústeres más cercanos se combinan iterativamente, formando una estructura de árbol o dendrograma que ilustra las relaciones jerárquicas entre los grupos. Este tratamiento no solo simplifica la identificación de patrones, sino que encima facilita la comprensión de la estructura subyacente de los datos, lo que resulta esencial para la toma de decisiones informadas.

La importancia de la AHC en el análisis de datos radica en su versatilidad y aplicabilidad. Consigue ser utilizada en una amplia gama de contextos, desde la segmentación de clientes en estudios de mercado hasta la clasificación de especies en ecología. Así, la AHC es particularmente valiosa en situaciones donde no se dispone de una etiqueta previa para los datos, permitiendo así el descubrimiento de grupos ocultos y la identificación de tendencias que de otro modo podrían pasar desapercibidas.

3.4.1 Fundamentos de la Agrupación Jerárquica Aglomerativa

La Agrupación Jerárquica Aglomerativa (AHC) es una técnica de análisis de datos que posibilita organizar un conjunto de objetos en grupos o clústeres de forma jerárquica. Este planteamiento es fundamental en el análisis exploratorio de datos, ya que facilita la identificación de patrones y relaciones dentro de los datos. La Agrupación Jerárquica Aglomerativa es un método de clustering que inicia con cada objeto como un clúster independiente y, a través de un proceso iterativo, fusiona los clústeres más cercanos hasta que se forma un único clúster que contiene todos los objetos (Taha, 2023). Esta técnica se caracteriza por su estructura jerárquica, que consigue visualizarse mediante un dendrograma, un diagrama que representa la agrupación de los objetos a diferentes niveles de similitud.

Entre las características más destacadas de la AHC se incluye su capacidad para dedicar una visión clara de la relación entre los datos, permitiendo al

analista seleccionar el número óptimo de clústeres según sus necesidades específicas. Al respecto, la AHC no requiere que se especifique previamente el número de clústeres, lo que la convierte en una herramienta flexible y exploratoria. Existen varios métodos para llevar a cabo la Agrupación Jerárquica Aglomerativa, cada uno de los cuales utiliza diferentes criterios para medir la distancia o similitud entre los clústeres. Los métodos más comunes incluyen:

i. *Método de enlace sencillo (Single Linkage):* Se basa en la distancia mínima entre los puntos de dos clústeres.

ii. *Método de enlace completo (Complete Linkage):* Utiliza la distancia máxima entre los puntos de los clústeres.

iii. *Método de enlace promedio (Average Linkage):* Calcula la distancia promedio entre todos los pares de puntos de los clústeres.

iv. *Método de Ward:* Minimiza la varianza dentro de los clústeres al fusionar los grupos, lo cual tiende a crear clústeres de tamaño similar.

Cada uno de estos métodos consigue influir en la forma en que los datos se agrupan, y la elección del método adecuado depende de la naturaleza de los datos y los objetivos del análisis. La AHC presenta varias ventajas que la hacen una opción atractiva para el análisis de datos:

- *No requiere suposiciones sobre la distribución de los datos:* A diferencia de otros métodos de agrupación, la AHC no asume que los datos siguen una distribución específica.

- *Visualización clara:* El dendrograma posibilita una interpretación visual de la relación entre los clústeres, facilitando la toma de decisiones.

- *Flexibilidad:* Posibilita la agrupación de datos de diferentes tipos y estructuras.

- *Sensibilidad a los valores atípicos:* Los outliers consiguen influir significativamente en la formación de clústeres, distorsionando los resultados.

- *Requerimiento de tiempo y recursos computacionales:* La AHC consigue ser computacionalmente costosa, especialmente con grandes conjuntos de datos, debido a la necesidad de calcular distancias entre todos los pares de objetos.

- *Dificultad en la selección del número óptimo de clústeres:* Aunque el dendrograma proporciona información valiosa, la elección del punto de corte para determinar el número de clústeres consigue ser subjetiva.

A fin de cuentas, la Agrupación Jerárquica Aglomerativa es una técnica poderosa y versátil para el análisis de datos, con características y métodos que la convierten en una herramienta fundamental para el descubrimiento de patrones y estructuras en conjuntos de datos complejos.

3.4.2 Método de Clasificación de Componentes Principales (MCA)

El Método de Clasificación de Componentes Principales (MCA, por sus siglas en inglés) es una técnica estadística utilizada para la reducción de la dimensionalidad y el análisis de datos categóricos. Su objetivo principal es transformar un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables, denominadas componentes principales, que son ortogonales entre sí y retienen la mayor parte de la variabilidad presente en los datos originales (Williams, 2002). A través de este proceso, se facilita la visualización y el análisis de datos complejos, permitiendo identificar patrones y relaciones que consiguen no ser evidentes en los datos originales.

El MCA es especialmente útil en situaciones donde se trabaja con variables cualitativas, como encuestas y estudios de mercado, donde se busca entender las relaciones entre diferentes categorías. Al aplicar el MCA, los analistas consiguen reducir el número de dimensiones y, al mismo tiempo, conservar la información más relevante, lo que resulta en un conjunto de datos más manejable y comprensible. La implementación del MCA implica varios pasos fundamentales:

i. *Recolección y preparación de datos:* Es crucial contar con un conjunto de datos bien estructurado. Esto incluye la identificación de las variables categóricas relevantes y la codificación adecuada de los datos.

ii. *Cálculo de la matriz de correspondencia:* A partir de los datos, se construye una matriz que refleja las frecuencias de las categorías para cada variable. Esta matriz es esencial para el siguiente paso.

iii. *Análisis de la matriz:* Se lleva a cabo un análisis de la matriz de correspondencia utilizando técnicas algebraicas que posibilitan extraer los componentes principales. Esto implica la descomposición en valores singulares o la realización de un análisis factorial.

iv. *Interpretación de los componentes:* Una vez obtenidos los componentes, es fundamental interpretarlos. Esto incluye analizar las cargas de cada variable en los componentes y determinar qué significan en el contexto del estudio.

v. *Visualización:* Se consiguen crear gráficos, como mapas de calor o diagramas de dispersión, que posibilitan visualizar las relaciones entre las variables y los componentes, facilitando la interpretación de los resultados.

La relación entre el Método de Clasificación de Componentes Principales (MCA) y la Agrupación Jerárquica Aglomerativa (AHC) es fundamental en el análisis de datos. El MCA se utiliza frecuentemente como un paso previo a la aplicación de AHC, ya que la reducción de dimensionalidad que proporciona el MCA posibilita que la agrupación de datos sea más efectiva y eficiente.

Una vez que se han obtenido los componentes principales a través del MCA, estos consiguen ser utilizados como nuevas variables en el análisis de agrupación. Esto no solo simplifica el proceso, sino que todavía mejora la calidad de los resultados, ya que se trabaja con datos que reflejan las relaciones más significativas. La combinación de estas dos técnicas posibilita a los analistas identificar grupos o clústeres dentro de los datos que tienen características similares, lo que es invaluable para la toma de decisiones informadas y la generación de percepciones fundamentadas en diversos campos, como el marketing, la biología y la psicología, entre otros.

La ejecución de una agrupación jerárquica aglomerativa (AHC) tras la aplicación de un análisis de componentes principales (MCA) implica una serie de pasos metódicos que garantizan que los datos estén preparados y analizados de manera efectiva. Antes de proceder a la ejecución de AHC, es crucial que los datos resultantes del MCA sean adecuadamente preparados, esto incluye:

i. *Selección de componentes:* Tras realizar el MCA, es necesario seleccionar un número adecuado de componentes que expliquen la mayor varianza posible en los datos. Generalmente, se eligen aquellos componentes que juntos explican al menos el 70-80% de la varianza total.

ii. *Normalización de datos:* Dado que AHC es sensible a la escala de los datos, es fundamental normalizar los componentes seleccionados. Esto se consigue lograr mediante la estandarización (restar la media y dividir por la desviación estándar) o mediante la normalización min-max.

iii. *Creación de la matriz de distancias:* Con los datos normalizados, se procede a calcular la matriz de distancias que se utilizará durante el proceso de agrupación. Existen varias métricas de distancia que se consiguen considerar, como la distancia euclidiana o la distancia de Manhattan, dependiendo de la naturaleza de los datos y los objetivos del análisis.

Una vez que los datos están preparados, se consigue proceder a la implementación del algoritmo de AHC. Este proceso involucra varios pasos:

i. *Elección del método de enlace:* AHC consigue utilizar diferentes métodos de enlace que determinan cómo se agrupan los clústeres. Los métodos más comunes son el enlace completo, el enlace simple y el enlace promedio. La elección del método consigue influir en la forma en que se forman los grupos.

ii. *Construcción del dendrograma:* El dendrograma es una representación gráfica que muestra la jerarquía de agrupación. A medida que se ejecuta el algoritmo, se van creando y uniendo clústeres, y el dendrograma refleja estas uniones, permitiendo visualizar cómo se forman los grupos a diferentes niveles de similitud.

iii. *Corte del dendrograma:* Una vez que se ha construido el dendrograma, es necesario decidir en qué nivel se realizará el corte para determinar el número final de grupos. Este corte consigue basarse en criterios visuales o en la identificación de un umbral de distancia que se considere adecuado para los objetivos del análisis.

La interpretación de los resultados de la AHC es fundamental para extraer conclusiones significativas del análisis. Algunos aspectos clave a considerar incluyen:

i. *Descripción de los grupos:* Cada grupo resultante de la AHC debe ser descrito en función de las características de los datos originales y los componentes seleccionados. Esto consigue incluir la media y la varianza de los valores dentro de cada grupo.

ii. *Identificación de patrones:* La agrupación consigue revelar patrones o tendencias interesantes que no eran evidentes en el análisis inicial. Es importante explorar las características que definen cada grupo y considerar su relevancia en el contexto del problema de investigación.

iii. *Validación de los resultados:* Para asegurar la robustez de los resultados obtenidos, se consiguen aplicar métodos de validación cruzada o comparar los grupos formados con otras técnicas de agrupación. Esto proporcionará una mayor confianza en las conclusiones obtenidas.

Como se ha dicho, la ejecución práctica de AHC tras un MCA es un proceso que requiere una preparación cuidadosa de los datos, una implementación metódica del algoritmo y una interpretación crítica de los resultados. Al seguir estos pasos, los investigadores consiguen aprovechar al máximo las técnicas de agrupación para descubrir información valiosa en sus conjuntos de datos.

Para Miyamoto (2022), la ejecución de la Agrupación Jerárquica Aglomerativa (AHC) tras la aplicación del Método de Clasificación de Componentes Principales (MCA) se revela como una estrategia poderosa en el análisis de datos. Por ende, la AHC posibilita identificar estructuras y patrones en los datos que consiguen no ser evidentes a simple vista. Al agrupar observaciones similares, se facilita la visualización y el análisis de la heterogeneidad dentro de los datos. Cuando se aplica después de un MCA, los resultados son aún más enriquecedores, ya que el MCA reduce la dimensionalidad de los datos, eliminando ruido y redundancias, y destacando las variaciones más significativas. Esto no solo optimiza el proceso de agrupación, sino que en esa misma línea mejora la calidad de las conclusiones obtenidas.

Incluso, la relación entre el MCA y la AHC enfatiza la importancia de la preparación de datos. Un MCA bien ejecutado proporciona un conjunto de variables que capta la esencia de los datos originales, lo que posibilita a la AHC trabajar con información más relevante y representativa. Esta perspectiva combinada maximiza la interpretabilidad de los resultados, permitiendo a los analistas realizar inferencias más acertadas sobre las relaciones y patrones que subyacen en los datos.

Por otro lado, es importante reconocer que, si bien la combinación de estas técnicas es poderosa, también presenta desafíos. La elección de los métodos de agrupación, la determinación del número de clústeres y la interpretación de los resultados requieren un juicio cuidadoso y una comprensión clara de los

objetivos del análisis. Los analistas deben estar atentos a las limitaciones inherentes a cada método y ser críticos en la interpretación de los resultados

La ejecución de AHC tras un MCA no solo es relevante, sino que se ha establecido como una práctica esencial en el análisis de datos modernos. Su capacidad para desglosar y comprender la complejidad de grandes volúmenes de datos la convierte en una herramienta valiosa para investigadores y profesionales en múltiples disciplinas. A medida que la cantidad de datos disponibles continúa creciendo, la integración de técnicas como la AHC y el MCA será fundamental para extraer conocimiento significativo y aplicable en la toma de decisiones.

Capítulo IV

Aprendizaje automático

4.1 Configurar y entrenar un clasificador XGBOOST: Indicadores de rendimiento de los modelos de predicción

XGBoost, que significa Extreme Gradient Boosting, es una biblioteca de código abierto que se ha convertido en un referente en el campo del aprendizaje automático. Desde su creación, ha sido adoptada extensamente por científicos de datos y profesionales de la inteligencia artificial, gracias a su capacidad para manejar grandes volúmenes de datos y su notable eficiencia en la ejecución de algoritmos de aprendizaje. XGBoost se basa en el concepto de boosting, una técnica que combina múltiples modelos débiles para crear un modelo fuerte que mejora la precisión en las predicciones (Wade y Glynn, 2020).

Entre las características más destacadas de XGBoost es su velocidad y rendimiento, lo que lo convierte en una herramienta ideal para competiciones de machine learning y aplicaciones en el mundo real donde el tiempo de respuesta es crítico. A través de optimizaciones específicas, como el uso de estructuras de datos en forma de árbol y el paralelismo en el proceso de entrenamiento, XGBoost consigue procesar datos de manera más rápida y efectiva que otras bibliotecas de aprendizaje automático.

De igual manera de su velocidad, XGBoost ofrece una variedad de funcionalidades que lo hacen extremadamente versátil. Posibilita manejar tanto problemas de clasificación como de regresión, y su capacidad para trabajar con datos faltantes y su robustez ante el sobreajuste son características que lo distinguen de otros algoritmos. También proporciona herramientas para la interpretación de modelos, lo que facilita entender cómo toman decisiones, un aspecto crucial en entornos donde la transparencia es fundamental.

La comunidad activa que rodea a XGBoost ha contribuido a su popularidad, pues, con una amplia documentación, tutoriales y ejemplos de uso, los usuarios consiguen aprender y aplicar XGBoost a sus propios problemas de manera efectiva. Además, su integración con otros frameworks de aprendizaje automático, como scikit-learn y TensorFlow, posibilita una mayor flexibilidad y

facilidad de uso en proyectos de ciencia de datos. En suma, XGBoost no solo destaca por su rendimiento superior y rápida ejecución, sino encima por su versatilidad y capacidad de interpretación. En un mundo donde las decisiones impulsadas por datos son cada vez más importantes, comprender y aplicar XGBoost se ha vuelto esencial para quienes buscan desarrollar modelos de predicción eficaces y eficientes en el ámbito del aprendizaje automático.

La instalación y configuración adecuada del entorno para trabajar con XGBoost es un paso crucial para garantizar que podamos aprovechar al máximo las capacidades de esta poderosa biblioteca de aprendizaje automático. Antes de proceder con la instalación de XGBoost, es importante asegurarse de que nuestro sistema cumpla con ciertos requisitos previos. A continuación, se enumeran los componentes más relevantes que necesitamos tener instalados:

i. *Python:* XGBoost es compatible con Python, por lo que se recomienda tener instalada una versión de Python iii.6 o superior. Consigue descargarse desde [python.org](https://www.python.org/downloads/).

ii. *Pip:* Este es el gestor de paquetes de Python y generalmente se instala automáticamente con Python. Asegúrate de tener pip actualizado ejecutando el siguiente comando en tu terminal:

bash

pip install --upgrade pip

iii. *Bibliotecas adicionales:* Para trabajar con datos y realizar análisis, es recomendable tener instaladas bibliotecas como numpy, pandas y scikit-learn.

Una vez que hemos verificado que contamos con los requisitos previos, el siguiente paso es instalar XGBoost. Este proceso es bastante sencillo y se consigue realizar a través de pip. Ejecuta el siguiente comando en tu terminal:

bash

pip install xgboost

Este comando ilustrará la última versión de XGBoost disponible en el repositorio de PyPI. Al igual consigues instalar una versión específica de XGBoost si es necesario, utilizando el siguiente formato:

bash

```
pip install xgboost==<version>
```

Después de instalar XGBoost, es recomendable configurar un entorno de desarrollo adecuado para facilitar nuestro trabajo. Dependiendo de tus preferencias, consigues elegir entre diversas herramientas, como Jupyter Notebook, PyCharm o Visual Studio Code. A continuación, se describen brevemente dos de las opciones más populares:

i. *Jupyter Notebook:* Esta herramienta posibilita crear y compartir documentos que contienen código, ecuaciones y visualizaciones. Para instalar Jupyter Notebook, consigues usar pip:

```
bash
pip install notebook
```

Una vez instalado, consigues iniciar Jupyter Notebook ejecutando el siguiente comando en tu terminal:

```
bash
jupyter notebook
```

ii. *Visual Studio Code:* Este editor de código ligero es altamente configurable y soporta una amplia gama de extensiones. Para trabajar con Python, asegúrate de instalar la extensión de Python en Visual Studio Code. Consigues descargarlo desde code.visualstudio.com.

Con la instalación de XGBoost y la configuración de tu entorno de desarrollo completadas, estarás listo para comenzar a entrenar modelos de predicción utilizando esta potente herramienta. El entrenamiento de un clasificador XGBoost es un proceso que implica varias etapas clave, desde la preparación de los datos hasta la selección y ajuste de los hiperparámetros. La preparación de los datos es una fase crítica en el entrenamiento de modelos de aprendizaje automático. Para XGBoost, es esencial que los datos estén en un formato adecuado y que se realicen algunas transformaciones y limpiezas previas (van Maarseveen, 2023). Los pasos típicos incluyen:

i. *Limpieza de datos:* Eliminar o imputar valores faltantes y corregir datos inconsistentes.

ii. *Codificación de variables categóricas:* Convertir variables categóricas en un formato numérico utilizando técnicas como one-hot encoding o label encoding.

iii. *Normalización y escalado:* Ajustar las características para que tengan una escala similar, lo cual consigue mejorar el rendimiento del modelo.

iv. *División de los datos:* Separar el conjunto de datos en conjuntos de entrenamiento y prueba, asegurando que el modelo se evalúe en datos no vistos.

XGBoost cuenta con una variedad de parámetros que consiguen ser ajustados para optimizar el rendimiento del modelo. Algunos de los más importantes incluyen:

- *learning_rate (tasa de aprendizaje):* Controla el impacto de cada árbol en el modelo final. Un valor más bajo consigue resultar en un mejor ajuste, pero requerirá más árboles para alcanzar la misma precisión.

- *n_estimators (número de árboles):* Define cuántos árboles se construirán. Un número mayor consigue mejorar la precisión, pero igualmente aumenta el riesgo de sobreajuste.

- *max_depth (profundidad máxima):* Limita la profundidad de cada árbol, lo que consigue ayudar a controlar la complejidad del modelo.

- *subsample (submuestreo):* Proporción de muestras utilizadas para entrenar cada árbol, que consigue ayudar a prevenir el sobreajuste si se establece en un valor menor a i.

Para encontrar la combinación óptima de estos hiperparámetros, se recomienda utilizar técnicas como la búsqueda en cuadrícula (grid search) o la búsqueda aleatoria (random search). Estas técnicas posibilitan evaluar múltiples combinaciones de parámetros y seleccionar la que ofrezca el mejor rendimiento en un conjunto de validación. Para ilustrar el proceso de entrenamiento de un clasificador XGBoost, consideremos un conjunto de datos ficticio de clasificación binaria. Supongamos que hemos preparado nuestros datos y dividido en conjuntos de entrenamiento y validación.

```
python

import xgboost as xgb

from sklearn.model_selection import train_test_split
```

```
from sklearn.metrics import accuracy_score
```

Supongamos que X es nuestro conjunto de características y y son las etiquetas

```
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)
```

Convertimos los datos a formato DMatrix de XGBoost

```
dtrain = xgb.DMatrix(X_train, label=y_train)
dval = xgb.DMatrix(X_val, label=y_val)
```

Definimos los parámetros del modelo

```
params = {
    'objective': 'binary:logistic',
    'learning_rate': 0.1,
    'max_depth': 5,
    'n_estimators': 100,
    'eval_metric': 'logloss'
}
```

Entrenamos el modelo

```
model = xgb.train(params, dtrain, num_boost_round=100)
```

Realizamos predicciones en el conjunto de validación

```
predictions = model.predict(dval)
predictions_binary = [1 if pred > 0.5 else 0 for pred in predictions]
```

Evaluamos el rendimiento del modelo

```
accuracy = accuracy_score(y_val, predictions_binary)
print(f'Precisión del modelo: {accuracy:.2f}')
```

En este ejemplo, hemos creado un modelo de clasificación binaria utilizando XGBoost. Tras entrenar el modelo, realizamos predicciones en el conjunto de validación y calculamos la precisión del modelo como métrica de

rendimiento inicial. La evaluación del rendimiento de un modelo de predicción es una etapa crucial en el desarrollo de cualquier aplicación de aprendizaje automático. Posibilita no solo medir la efectividad del modelo en datos no vistos, sino también identificar áreas de mejora y optimización.

Al evaluar un clasificador, es fundamental elegir las métricas adecuadas que reflejen el rendimiento del modelo de manera precisa. Las tres métricas más comunes son la precisión, el recall y el F1-score:

i. *Precisión:* Se refiere a la proporción de verdaderos positivos sobre el total de predicciones positivas realizadas por el modelo. Es útil cuando el costo de las falsas alarmas es alto. Se calcula como:

\[

\text{Precisión} = \frac{TP}{TP + FP}

\]

donde \(TP\) son los verdaderos positivos y \(FP\) son los falsos positivos.

ii. *Recall:* Asimismo conocido como sensibilidad, mide la capacidad del modelo para identificar correctamente los casos positivos. Es especialmente importante en aplicaciones donde se desea minimizar los falsos negativos. Se calcula como:

\[

\text{Recall} = \frac{TP}{TP + FN}

\]

donde \(FN\) son los falsos negativos.

iii. *F1-score:* Esta métrica combina la precisión y el recall en un solo valor, proporcionando una medida balanceada que es especialmente útil en conjuntos de datos desbalanceados. Se calcula como:

\[

F1 = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}

\]

Un alto F1-score indica un buen equilibrio entre precisión y recall. Así de las métricas mencionadas, las curvas ROC (Receiver Operating Characteristic) y el AUC (Area Under the Curve) son herramientas valiosas para evaluar el rendimiento de un modelo de clasificación.

- *Curva ROC:* Esta curva traza la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) a diferentes umbrales de decisión. Una curva ROC que se aproxima a la esquina superior izquierda del gráfico indica un mejor rendimiento del modelo.

- *AUC:* El área bajo la curva ROC proporciona una medida única del rendimiento del modelo. Un AUC de 0.5 sugiere que el modelo no es mejor que un modelo aleatorio, mientras que un AUC de i.0 indica un modelo perfecto. Un AUC superior a 0.7 se considera generalmente aceptable, y valores superiores a 0.9 son indicativos de un modelo excelente.

La interpretación de los resultados obtenidos a partir de las métricas de rendimiento es fundamental para realizar ajustes finales en el modelo. Si la precisión es alta pero el recall es bajo, consigue ser un indicativo de que el modelo está clasificando correctamente los casos positivos pero está perdiendo muchos de ellos. En este caso, podría ser necesario ajustar el umbral de decisión o considerar técnicas de rebalancing de clases.

En suma, se consiguen utilizar métodos como la validación cruzada para obtener estimaciones más robustas del rendimiento del modelo y minimizar el riesgo de sobreajuste. La validación cruzada implica dividir el conjunto de datos en múltiples subconjuntos y entrenar el modelo en diferentes combinaciones de estos, permitiendo así evaluar su rendimiento en una variedad de configuraciones. Es recomendable realizar un análisis de error para comprender mejor qué tipos de instancias el modelo está clasificando incorrectamente. Esto consigue proporcionar información valiosa sobre cómo mejorar el modelo, ya sea mediante la recolección de más datos, la selección de nuevas características o el ajuste de hiperparámetros.

Así pues, la evaluación del rendimiento es un paso fundamental en el ciclo de vida del desarrollo de un modelo de aprendizaje automático. Utilizando una combinación de métricas de rendimiento y técnicas de visualización como las curvas ROC, es posible obtener una comprensión profunda de la efectividad del

clasificador XGBoost y realizar las mejoras necesarias para optimizar su desempeño.

XGBoost ha demostrado ser una herramienta poderosa y eficiente para la construcción de modelos predictivos, gracias a su capacidad para manejar grandes volúmenes de datos y su robustez contra el sobreajuste. A través del proceso de instalación, preparación de datos, ajuste de hiperparámetros y evaluación del rendimiento, hemos establecido un camino claro para los usuarios que deseen implementar esta técnica en sus proyectos (Chen y Guestrin, 2016). Algunas áreas a considerar para el futuro incluyen:

i. *Exploración de nuevas funcionalidades:* XGBoost sigue evolucionando, y nuevas versiones a menudo traen consigo mejoras en la eficiencia y la precisión. Mantenerse al tanto de la documentación oficial y las actualizaciones consigue ofrecer ventajas significativas en tus modelos.

ii. *Hiperparámetros avanzados:* El ajuste de hiperparámetros más complejos consigue llevar a un rendimiento aún mejor, experimentar con las opciones de regularización, tasas de aprendizaje adaptativas y el uso de técnicas como la búsqueda en cuadrícula o la optimización bayesiana consigue ser beneficioso.

iii. *Integración con otras herramientas:* XGBoost se consigue combinar con otras bibliotecas de aprendizaje automático y herramientas de análisis de datos, como Scikit-learn y pandas. Explorar estas integraciones consigue permitir un flujo de trabajo más eficiente y mejorar la colaboración entre diferentes modelos.

iv. *Interpretación y visualización:* La interpretación de los resultados de un modelo es crucial para la toma de decisiones informadas. Utilizar herramientas como SHAP (SHapley Additive exPlanations) o LIME (Local Interpretable Model-agnostic Explanations) consigue ayudar a desentrañar el funcionamiento interno de tu clasificador y a comunicar mejor sus decisiones a los stakeholders.

v. *Aplicaciones en problemas del mundo real:* XGBoost se ha utilizado con éxito en una amplia variedad de dominios, incluyendo finanzas, salud y marketing. Considerar la aplicación de XGBoost a nuevos problemas o conjuntos de datos consigue ampliar su utilidad y efectividad.

Dominar XGBoost no solo implica entender cómo entrenar un modelo, sino encima cómo adaptarlo y mejorarlo continuamente. Al seguir explorando nuevas técnicas y aplicaciones, los profesionales del aprendizaje automático

consiguen maximizar el potencial de esta potente herramienta y contribuir a soluciones innovadoras en diversos campos.

4.2 Configurar e interpretar una agrupación DBSCAN y una agrupación difusa k-means

La agrupación de datos, también conocida como clustering, es una técnica fundamental en el análisis de datos que tiene como objetivo agrupar un conjunto de objetos de tal manera que los objetos dentro de un mismo grupo (o clúster) sean más similares entre sí que aquellos que pertenecen a otros grupos. Esta técnica se utiliza ampliamente en diversas disciplinas, como la minería de datos, el aprendizaje automático y la inteligencia artificial, y posibilita descubrir patrones ocultos en conjuntos de datos complejos.

Existen múltiples perspectivas para llevar a cabo la agrupación, cada uno con sus propias características y aplicaciones. Los métodos de agrupación consiguen clasificarse en dos categorías principales: aquellos basados en la densidad, como DBSCAN, y los basados en centroides, como el k-means difuso. La elección del método adecuado depende de la naturaleza de los datos, la forma de los grupos esperados y los objetivos específicos del análisis. La agrupación no solo ayuda a organizar datos en categorías significativas, sino que también facilita la visualización y la interpretación de grandes volúmenes de información. A través de la agrupación, se consiguen identificar tendencias, anomalías y relaciones que de otro modo podrían pasar desapercibidas.

La técnica de agrupación DBSCAN (Density-Based Spatial Clustering of Applications with Noise) se ha convertido en una herramienta fundamental en el análisis de datos, especialmente en aquellos conjuntos que presentan formas y densidades no lineales. Su capacidad para identificar agrupamientos de alta densidad y separar el ruido de manera efectiva la hace ideal para una variedad de aplicaciones, desde la detección de anomalías hasta la segmentación de mercado.

DBSCAN es un algoritmo de agrupación basado en la densidad que clasifica puntos en diferentes grupos según la densidad de puntos en su vecindad. La premisa básica de DBSCAN es que un grupo se forma alrededor de áreas de alta densidad de puntos, mientras que los puntos que no pertenecen a ningún grupo se clasifican como ruido (Ashour y Sunoallah, 2011). Este método

se basa en dos conceptos clave: el epsilon (ε), que define la distancia máxima para considerar que dos puntos están en la misma vecindad, y el mínimo de puntos (MinPts), que indica cuántos puntos son necesarios para formar un núcleo denso.

Para configurar DBSCAN de manera efectiva, es crucial seleccionar adecuadamente dos parámetros:

i. *Epsilon (ε):* Este parámetro define la distancia máxima entre dos puntos para que se consideren parte de la misma agrupación. Un valor demasiado bajo consigue llevar a la formación de muchas agrupaciones pequeñas, mientras que un valor demasiado alto consigue resultar en una única agrupación que engloba todos los puntos.

ii. *Mínimo de puntos (MinPts):* Este parámetro establece el número mínimo de puntos requeridos para formar un núcleo. Generalmente, se recomienda que MinPts sea al menos igual al número de dimensiones del espacio de datos más uno. En un espacio de dos dimensiones, un valor de MinPts de 3 es un buen punto de partida.

La elección adecuada de estos parámetros consigue requerir iteraciones y experimentación, a menudo apoyándose en métodos como el gráfico de k-distancia para determinar un valor apropiado de ε. Una vez configurado y ejecutado el algoritmo, es fundamental interpretar los resultados obtenidos. DBSCAN categoriza los puntos en tres tipos: núcleos, borde y ruido. Los puntos de núcleo son aquellos que tienen al menos MinPts en su vecindad ε; los puntos de borde son aquellos que están dentro de la vecindad de un núcleo pero no cumplen con el criterio de MinPts; y el ruido son aquellos que no pertenecen a ninguna agrupación.

Al analizar los resultados, es importante observar la cantidad de agrupaciones formadas y la cantidad de puntos clasificados como ruido. Un alto porcentaje de ruido consigue indicar que la configuración de parámetros necesita ajustes, mientras que varios núcleos densos sugieren que el algoritmo ha capturado correctamente las estructuras subyacentes de los datos. La visualización de los resultados mediante gráficos de dispersión consigue ser especialmente útil para evaluar la calidad de las agrupaciones y para identificar patrones o outliers en los datos.

En otras palabras, la configuración y la interpretación de los resultados en DBSCAN son pasos cruciales que requieren atención a los detalles y una comprensión clara de los parámetros involucrados. Con una configuración adecuada, DBSCAN consigue revelar valiosa información sobre la estructura de los datos, facilitando así decisiones informadas en diversos campos de aplicación.

El k-means difuso es una variante del clásico algoritmo k-means que posibilita una asignación más flexible de los datos a los clústeres. En lugar de asignar cada punto de datos a un único clúster, el k-means difuso posibilita que un punto pertenezca a múltiples clústeres con diferentes grados de pertenencia. Esta característica es especialmente útil en situaciones donde los límites entre los grupos no son claros y los datos presentan solapamientos significativos.

El k-means difuso se basa en el concepto de partición difusa, que se diferencia de la partición crisp del k-means tradicional. En este planteamiento, cada punto de datos se asigna a un clúster de acuerdo con un grado de pertenencia que varía entre 0 y i (Valova et al., 2024). Este grado indica la fuerza de la pertenencia de un punto a un clúster específico; la suma de los grados de pertenencia de un punto a todos los clústeres es igual a i. Esto posibilita una representación más matizada de la estructura de los datos, donde los puntos cercanos a los límites de los clústeres consiguen ser considerados como parte de múltiples grupos. El k-means difuso tiene varios parámetros clave que deben ser configurados para obtener resultados óptimos. Los más relevantes son:

i. *Número de clústeres (k):* Al igual que en el k-means clásico, este parámetro determina cuántos grupos se intentará identificar en los datos. Una elección adecuada de k es crucial, ya que un valor demasiado bajo consigue resultar en una pérdida de información, mientras que un valor demasiado alto consigue llevar a un sobreajuste de los datos.

ii. *Coeficiente de difusividad (m):* Este parámetro controla la "difusión" de la pertenencia de los puntos a los clústeres. Un valor de m mayor que 1 hace que la pertenencia a los clústeres sea más difusa, permitiendo que los puntos se distribuyan más equitativamente entre los clústeres. Por otro lado, un valor de m igual a 1 convierte el algoritmo en un k-means clásico, donde cada punto pertenece exclusivamente a un único clúster.

iii. *Criterio de convergencia:* Este parámetro define cuándo el algoritmo debe detenerse. Comúnmente, se utiliza un umbral basado en la variación de los

centros de clústeres o en la suma de los errores de pertenencia de los puntos a los clústeres.

La interpretación de los resultados en k-means difuso implica analizar tanto los clústeres formados como los grados de pertenencia de cada punto. Los clústeres se consiguen visualizar mediante gráficos, donde cada punto es representado con un color que indica su pertenencia a los clústeres, y la intensidad de ese color refleja el grado de pertenencia. Al evaluar los resultados, es importante considerar no solo la configuración de los clústeres, sino al igual los grados de pertenencia. Un punto que tiene una alta pertenencia a múltiples clústeres consigue indicar solapamiento natural en los datos o la presencia de características diversas en ese punto.

Por lo tanto, al interpretar los resultados, se debe tener en cuenta esta flexibilidad que brinda el k-means difuso, así como su utilidad en la toma de decisiones informadas sobre la estructura de los datos. A fin de cuentas, la configuración adecuada del k-means difuso y la interpretación cuidadosa de sus resultados son fundamentales para aprovechar al máximo este método de agrupación en contextos donde la ambigüedad entre categorías es prevalente. Al abordar la agrupación de datos, tanto DBSCAN como k-means difuso son técnicas populares que presentan diferentes tratamientos y beneficios dependiendo del contexto de la aplicación.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) destaca por su capacidad para identificar grupos de datos de forma no euclidiana, lo que lo convierte en un método ideal para datos con estructuras complejas y distribuciones irregulares. Entre sus ventajas se incluyen:

- *Detección de ruido:* DBSCAN consigue manejar datos ruidosos y outliers sin afectar significativamente el resultado de la agrupación.

- *No requiere número de grupos predefinido:* A diferencia de k-means, DBSCAN no requiere que se especifique el número de grupos de antemano, lo que lo hace más flexible en escenarios desconocidos.

- *Agrupaciones de forma arbitraria:* Este método es capaz de detectar agrupaciones de formas no esféricas, lo que es especialmente útil en conjuntos de datos donde las agrupaciones no son homogéneas.

Empero, DBSCAN también tiene desventajas, como la sensibilidad a los parámetros de entrada, en particular el radio de búsqueda y el número mínimo de puntos necesarios para formar un grupo. Si estos parámetros no se ajustan adecuadamente, consigue resultar en agrupaciones erróneas o en la omisión de grupos significativos. Por otro lado, k-means difuso (o fuzzy k-means) ofrece una forma de asignar puntos a múltiples grupos con diferentes grados de pertenencia, lo que posibilita una mayor flexibilidad en la clasificación. Sus ventajas incluyen:

- *Asignación de pertenencia difusa:* Posibilita que un punto pertenezca a más de un grupo, lo que es útil en situaciones donde los límites entre grupos no son claros.

- *Simplicidad y rapidez:* El algoritmo es relativamente simple y se ejecuta rápidamente en comparación con métodos más complejos, lo que lo hace adecuado para conjuntos de datos grandes.

- *Facilidad de interpretación:* Los resultados son generalmente fáciles de entender, dado que se basa en la distancia euclidiana y la media de los grupos.

Ahora bien, k-means difuso tiene sus propias desventajas. Requiere que se especifique el número de grupos de antemano, lo cual consigue ser un desafío si no se tiene una idea clara de la estructura de los datos. Incluso, es sensible a la inicialización de los centros de los grupos, lo que consigue llevar a resultados inconsistentes si no se eligen adecuadamente. La elección entre DBSCAN y k-means difuso consigue depender en gran medida del tipo de datos y del objetivo del análisis. DBSCAN es particularmente efectivo en los siguientes escenarios:

- *Datos con ruido significativo:* En situaciones donde los outliers son comunes, como en la detección de fraudes o análisis de redes sociales, DBSCAN consigue proporcionar una agrupación más robusta.

- *Agrupaciones de forma arbitraria:* Cuando se trabaja con datos geoespaciales o datos de sensores, donde las agrupaciones consiguen ser de formas irregulares, DBSCAN es el método preferido.

Por otro lado, k-means difuso es recomendable en casos donde:

- *Los grupos son más homogéneos:* En situaciones donde se espera que los grupos sean más esféricos y bien definidos, como en la segmentación de mercado, k-means difuso consigue ser más eficiente.

- *La interpretación clara es clave:* En análisis donde la claridad en la pertenencia a grupos es crucial, como en la clasificación de imágenes o en estudios demográficos, este método consigue resultar más adecuado.

Al seleccionar entre DBSCAN y k-means difuso, es crucial considerar varios factores, incluyendo la naturaleza de los datos, la presencia de ruido, el número de grupos esperados y la necesidad de interpretabilidad. La experimentación con ambos métodos, así como la evaluación de sus resultados en el contexto específico de la aplicación, consigue proporcionar información valiosa y guiar en la elección del algoritmo más adecuado. Además, es recomendable realizar un análisis de sensibilidad sobre los parámetros utilizados para asegurar que los resultados sean consistentes y representativos de la estructura subyacente de los datos.

Tanto DBSCAN como k-means difuso presentan herramientas poderosas para la agrupación de datos, pero su efectividad dependerá del contexto y de cómo se configuren sus parámetros. La comprensión de sus diferencias y similitudes es fundamental para aprovechar al máximo estas técnicas en la práctica. DBSCAN, con su capacidad para identificar grupos de diferentes formas y tamaños, se destaca especialmente en situaciones donde los datos presentan ruido y distribuciones irregulares (Hahsler et al., 2019). Su planteamiento basado en la densidad posibilita una agrupación más flexible, lo que resulta en una interpretación más natural de los datos en muchas aplicaciones del mundo real.

Ahora bien, su dependencia de la elección de parámetros como el número mínimo de puntos y la distancia epsilon consigue ser un desafío, lo que requiere una cuidadosa calibración para obtener resultados óptimos. Por otro lado, el k-means difuso ofrece una alternativa robusta en contextos donde la pertenencia a los grupos no es estrictamente binaria. Al permitir que un punto de datos pertenezca a múltiples grupos con diferentes grados de pertenencia, este método es particularmente útil en situaciones donde los límites entre las categorías no son claros. Empero, su eficacia depende en gran medida de la selección del número de clústeres y consigue ser sensible a la inicialización de los centroides, lo que consigue llevar a resultados inconsistentes si no se maneja adecuadamente.

Al comparar ambos métodos, es evidente que la elección entre DBSCAN y k-means difuso debe basarse en la naturaleza del conjunto de datos y los objetivos específicos del análisis. DBSCAN es más adecuado para conjuntos de

datos con ruido y formas no esféricas, mientras que el k-means difuso consigue ser preferible en contextos donde la frontera entre las clases es difusa.

Tanto DBSCAN como k-means difuso son herramientas valiosas en el arsenal del analista de datos. La comprensión de sus principios, configuraciones y resultados interpretativos posibilita a los investigadores y profesionales tomar decisiones informadas que optimicen el proceso de agrupación y, en última instancia, contribuyan a obtener conocimientos más profundos y útiles de los datos analizados. Al final, la selección del método adecuado no solo mejora la calidad de los resultados obtenidos, sino que incluso potencia la capacidad de generar valor a partir de la información disponible.

4.3 Entrenamiento de una máquina de vectores soporte para regresión (SVR)

El aprendizaje automático ha revolucionado la forma en que abordamos problemas complejos en diversas áreas, desde la economía hasta la biomedicina. Dentro de este vasto campo, el soporte vectorial para regresión (SVR, por sus siglas en inglés) se destaca como una técnica poderosa y versátil para modelar relaciones no lineales entre variables.

El soporte vectorial para regresión es un método de aprendizaje automático que se basa en la teoría de soporte vectorial, desarrollada inicialmente para tareas de clasificación. A diferencia de la clasificación, donde el objetivo es categorizar datos en diferentes clases, el SVR se utiliza para predecir valores continuos. El modelo busca encontrar una función que se ajuste a los datos con un margen de tolerancia especificado, permitiendo así que algunas desviaciones sean aceptables. Esta capacidad para manejar ruido y variaciones en los datos hace que el SVR sea especialmente útil en situaciones donde la precisión es crucial.

El concepto de soporte vectorial fue introducido por Vladimir Vapnik y su equipo en la década de 1990, inicialmente como una técnica para clasificación binaria. Pese a, su adaptación para la regresión surgió poco después, en 1997, cuando Vapnik y Alexey Chervonenkis publicaron su trabajo sobre SVR. Desde entonces, el SVR ha evolucionado significativamente, con diversas mejoras en algoritmos y técnicas que han ampliado su aplicabilidad (Rodríguez y Bajorath, 2022). En el contexto histórico, se han desarrollado variantes del SVR, como el

SVR basado en funciones de kernel, que posibilita capturar relaciones más complejas en los datos.

El SVR ha encontrado aplicaciones en múltiples dominios. En finanzas, se utiliza para predecir precios de acciones y tendencias del mercado. En el ámbito de la salud, se emplea para estimar la progresión de enfermedades y resultados de tratamientos. De igual manera, en la ingeniería, el SVR se ha utilizado para modelar procesos físicos y optimizar diseños. Su capacidad para manejar datos de alta dimensionalidad y su robustez ante el sobreajuste lo convierten en una herramienta valiosa en la caja de herramientas de los científicos de datos. Con esta introducción al SVR, se establece el contexto necesario para profundizar en los fundamentos teóricos que sustentan este método, lo que permitirá entender mejor su proceso de entrenamiento y las perspectivas futuras en su aplicación.

El entrenamiento de una máquina de vectores soporte para regresión (SVR) se basa en principios teóricos que son fundamentales para comprender su funcionamiento y efectividad. El aprendizaje automático es un subcampo de la inteligencia artificial que se dirige en el desarrollo de algoritmos que posibilitan a las máquinas aprender a partir de datos. Este proceso implica el uso de modelos matemáticos y estadísticos para identificar patrones en los datos y hacer predicciones o tomar decisiones basadas en ellos. Dentro del aprendizaje automático, existen diferentes planteamientos, como el aprendizaje supervisado, no supervisado y por refuerzo. El SVR se enmarca dentro del aprendizaje supervisado, donde se dispone de un conjunto de datos etiquetados que se utilizan para entrenar el modelo.

En el contexto del SVR, el objetivo es encontrar una función que se ajuste a los datos de entrenamiento de manera que pueda predecir los valores de salida de nuevas instancias. A diferencia de otros métodos de regresión que buscan minimizar el error cuadrático, el SVR tiene un planteamiento diferente que se centra en ajustar un margen alrededor de la función de regresión, lo que lo hace robusto frente a datos atípicos.

Uno de los conceptos clave en SVR es el margen, que se refiere al espacio entre la función de regresión y los puntos de datos más cercanos. A diferencia de los métodos de regresión tradicionales que buscan minimizar el error total, el SVR establece un margen dentro del cual se posibilita que los errores ocurran. Esto se traduce en una función de pérdida que no penaliza cada error de manera

cuadrática, sino que se enfoca en errores que superan un cierto umbral, conocido como epsilon (ε).

La función de pérdida utilizada en SVR es la función de pérdida ε-insensitive, que establece que los errores dentro de este margen no tienen penalización. Solo aquellos errores que superan este margen se penalizan, lo que posibilita que el modelo ignore pequeñas variaciones en los datos, contribuyendo a su robustez. Este planteamiento ayuda a evitar el sobreajuste y posibilita al modelo generalizar mejor a datos no vistos. El uso de funciones kernel es otro aspecto fundamental en el funcionamiento del SVR, por ende, un kernel es una función que posibilita transformar los datos de entrada en un espacio de características de mayor dimensión, donde se busca que se vuelvan linealmente separables (Mnyanghwalo et al., 2020). Este proceso se conoce como "mapeo implícito" y posibilita al SVR manejar de manera efectiva problemas no lineales.

Existen diferentes tipos de funciones kernel, como el kernel lineal, el polinómico y el radial (RBF), cada uno de los cuales tiene sus propias características y aplicaciones. La elección del kernel adecuado es crucial, ya que influye directamente en la capacidad del modelo para capturar la complejidad de los datos. Ahora bien, la selección de parámetros asociados al kernel, como el coeficiente de regularización y los parámetros del kernel en sí, al igual juega un papel importante en el rendimiento del modelo.

Dicho de otro modo, los fundamentos teóricos del SVR se basan en conceptos clave del aprendizaje automático, el manejo del margen y la función de pérdida, así como en la aplicación de funciones kernel para abordar problemas complejos. Estos elementos constituyen la base sobre la cual se construye el proceso de entrenamiento y se define la eficacia del modelo en la predicción de resultados en diversos contextos. El entrenamiento de una máquina de vectores soporte para regresión (SVR) implica una serie de pasos críticos que garantizan que el modelo sea capaz de hacer predicciones precisas y generalizables. Antes de comenzar el entrenamiento de un modelo SVR, es fundamental preparar y preprocesar los datos (Drucker et al., 1997). Esto incluye varias tareas:

i. *Recolección de datos:* Obtener un conjunto de datos representativo que contenga las variables independientes (características) y la variable dependiente (objetivo) para las cuales se desea realizar la regresión.

ii. *Limpieza de datos:* Identificar y manejar valores faltantes, errores o inconsistencias dentro del conjunto de datos. Esto consigue implicar la eliminación de registros incompletos o la imputación de valores donde sea necesario.

iii. *Normalización y estandarización:* Dado que el SVR es sensible a la escala de los datos, es importante normalizar o estandarizar las características. La normalización consigue implicar escalar los datos a un rango específico (entre 0 y 1), mientras que la estandarización ajusta los datos para que tengan una media de cero y una desviación estándar de uno.

iv. *División del conjunto de datos:* Separar el conjunto de datos en conjuntos de entrenamiento y prueba (y, en algunos casos, un conjunto de validación) es crucial para evaluar la capacidad de generalización del modelo. Una división típica podría ser un 70% para el entrenamiento y un 30% para la prueba.

Una vez que los datos han sido preparados, el siguiente paso es seleccionar los parámetros del modelo SVR. Esto incluye:

i. *Elección del tipo de kernel:* El SVR utiliza diferentes funciones de kernel que consiguen transformar los datos de entrada en un espacio de mayor dimensión, lo que posibilita al modelo capturar relaciones no lineales. Entre los kernels más comunes se encuentran el kernel lineal, el kernel polinómico y el kernel radial (RBF). La elección del kernel consigue afectar significativamente el rendimiento del modelo.

ii. *Ajuste de hiperparámetros:* El SVR tiene varios hiperparámetros que deben ajustarse, como el parámetro de regularización (C), que controla el equilibrio entre la complejidad del modelo y el margen de tolerancia en los errores, y el parámetro de epsilon (ε), que define el ancho de la franja dentro de la cual se posibilita el error. Técnicas como la validación cruzada son comunes para encontrar la combinación óptima de hiperparámetros.

iii. *Entrenamiento del modelo:* Con los parámetros seleccionados, se procede a entrenar el modelo SVR utilizando el conjunto de datos de entrenamiento. El algoritmo ajusta iterativamente los parámetros del modelo para minimizar la función de pérdida, que, en el caso del SVR, es a menudo una función de pérdida de ε-insensibilidad.

Una vez completado el entrenamiento del modelo SVR, es crucial evaluar su rendimiento utilizando el conjunto de prueba. Aquí, se emplean varias métricas para medir la precisión del modelo:

i. *Error cuadrático medio (MSE):* Esta métrica calcula la media de los cuadrados de los errores, lo que proporciona una medida clara de cuánto se desvían las predicciones del modelo de los valores reales.

ii. R^2 (coeficiente de determinación): Esta métrica evalúa qué tan bien se ajusta el modelo a los datos, proporcionando una indicación de la proporción de la variabilidad total que es explicada por el modelo.

iii. *Análisis de residuos:* Analizar los residuos (diferencia entre valores reales y predicciones) posibilita identificar patrones que el modelo consigue no estar capturando adecuadamente. Esto consigue llevar a ajustes adicionales en la preparación de datos o en los parámetros del modelo.

El éxito en el proceso de entrenamiento de un SVR depende en gran medida de la cuidadosa atención a cada uno de estos pasos. Un modelo bien entrenado no solo dedicará buenas predicciones en el conjunto de prueba, sino que todavía será capaz de generalizar a nuevos datos en situaciones prácticas. El proceso de entrenamiento de un modelo SVR implica varias etapas críticas, desde la preparación y el preprocesamiento de datos hasta la selección de parámetros y la evaluación del modelo. Cada uno de estos pasos es vital para garantizar que el modelo sea capaz de generalizar bien a nuevos datos y presentar predicciones precisas.

Uno de los principales problemas es la elección del kernel adecuado y su parametrización, ya que un kernel mal seleccionado consigue resultar en un rendimiento subóptimo. En espacial, el ajuste de los parámetros de regularización y el costo de entrenamiento consiguen ser complejos, especialmente en conjuntos de datos de gran dimensión. Otro desafío es la escalabilidad de SVR en comparación con otros algoritmos de regresión más simples, lo que consigue limitar su aplicación en escenarios donde se requiere un procesamiento en tiempo real.

Las investigaciones actuales se centran en la optimización de algoritmos y en el desarrollo de nuevas técnicas que permitan mejorar la eficiencia y la precisión de los modelos SVR. Entre las tendencias emergentes se encuentran el

uso de técnicas de aprendizaje profundo para complementar los métodos tradicionales de SVR, así como la integración de tratamientos de aprendizaje federado, que posibilitan construir modelos a partir de datos distribuidos sin necesidad de centralizarlos.

Asimismo, la incorporación de técnicas de interpretación de modelos es cada vez más relevante, ya que los usuarios buscan no solo predicciones precisas, sino del mismo modo una comprensión clara de cómo se toman estas decisiones. Esto consigue abrir nuevas oportunidades para el uso de SVR en sectores donde la transparencia y la interpretabilidad son fundamentales, como en la medicina y las finanzas. En definitiva, el SVR sigue siendo una herramienta valiosa en el arsenal del aprendizaje automático, y su desarrollo futuro promete ofrecer soluciones innovadoras y efectivas para enfrentar los desafíos actuales en la predicción y el análisis de datos.

4.4 Conjunto de datos para la clasificación: K Nearest Neighbors vs. Naive Bayes

La clasificación es una tarea fundamental en el campo del aprendizaje automático, donde el objetivo es asignar una etiqueta a una entrada desconocida basada en ejemplos previos. Entre las diversas técnicas de clasificación, dos de las más utilizadas son K Nearest Neighbors (KNN) y Naive Bayes. Estas técnicas son particularmente valoradas por su simplicidad y eficacia en una amplia gama de aplicaciones (Kramer, 2013).

K Nearest Neighbors es un algoritmo de clasificación basado en la cercanía entre los puntos de datos. Su funcionamiento se basa en la idea de que, si un nuevo dato se asemeja a los datos de entrenamiento en su proximidad, es probable que pertenezca a la misma clase. Para clasificar un nuevo elemento, KNN calcula la distancia entre este y todos los puntos de datos en el conjunto de entrenamiento, selecciona los K más cercanos y asigna la clase que más predominancia tenga entre estos vecinos. La elección del valor de K es crucial, ya que un K muy pequeño consigue hacer que el modelo sea sensible al ruido, mientras que un K muy grande consigue llevar a una clasificación más generalizada pero menos precisa.

Naive Bayes es una familia de algoritmos de clasificación que se basa en el teorema de Bayes y asume que las características del conjunto de datos son

independientes entre sí, lo que se conoce como la "naive" o ingenua suposición. Esta técnica es especialmente eficaz en problemas de clasificación de texto, como el filtrado de spam y la clasificación de documentos, donde las características (como la presencia o ausencia de ciertas palabras) son a menudo independientes. Naive Bayes calcula la probabilidad de que un dato pertenezca a cada clase y selecciona la clase con la mayor probabilidad como resultado de la clasificación.

La calidad y la adecuación de los conjuntos de datos son fundamentales para el rendimiento de cualquier algoritmo de clasificación. Un conjunto de datos bien diseñado y representativo asegura que el modelo pueda aprender patrones significativos y generalizables. En especial, la diversidad en las características y la cantidad de datos disponibles consigue influir en la capacidad del modelo para hacer predicciones precisas. Por lo tanto, es esencial prestar atención al diseño y la preparación de los conjuntos de datos antes de aplicar cualquier técnica de clasificación, ya que esto consigue determinar en gran medida la eficacia de KNN, Naive Bayes y otros algoritmos de aprendizaje automático.

Los conjuntos de datos son fundamentales en la implementación de algoritmos de clasificación como K Nearest Neighbors (KNN) y Naive Bayes. La calidad y las características de estos conjuntos consiguen influir significativamente en la precisión y efectividad de los modelos de clasificación. Los conjuntos de datos consiguen contener diferentes tipos de datos, que se clasifican principalmente en numéricos y categóricos (Madariaga et al., 2022). Los datos numéricos son aquellos que se consiguen medir y expresar en términos cuantitativos, como la altura, el peso o la temperatura. Por otro lado, los datos categóricos se refieren a variables que representan categorías o grupos, como el género, la raza o el tipo de producto.

KNN es un algoritmo que consigue manejar tanto datos numéricos como categóricos, pero consigue requerir transformaciones, como la codificación one-hot, para trabajar adecuadamente con variables categóricas. Naive Bayes, en cambio, se basa en una suposición de independencia entre las características y, si bien incluso consigue trabajar con ambos tipos de datos, sus variantes consiguen variar en complejidad dependiendo de la naturaleza de los datos.

El tamaño del conjunto de datos es otro factor crucial que consigue afectar el rendimiento de los modelos de clasificación. Un conjunto de datos más grande generalmente proporciona más información y ayuda a crear modelos más

robustos y generalizables. Pero, un tamaño excesivo también consigue presentar dilemas, como tiempos de procesamiento más largos y la necesidad de más recursos computacionales. La calidad del conjunto de datos es igualmente importante. Esto incluye la precisión de los datos, la presencia de valores faltantes, y la existencia de ruido o errores en las mediciones. Un conjunto de datos de alta calidad consigue mejorar significativamente la eficacia de la clasificación, mientras que un conjunto de datos de baja calidad consigue llevar a resultados engañosos y decisiones erróneas.

Antes de aplicar cualquier algoritmo de clasificación, es esencial realizar un preprocesamiento de los datos. Este proceso consigue incluir varias etapas, como la limpieza de datos, la normalización o estandarización de características numéricas, y la conversión de datos categóricos en un formato utilizable. La limpieza de datos implica la identificación y corrección de errores, así como la gestión de valores faltantes, que consiguen ser eliminados o imputados según las necesidades del análisis. La normalización y estandarización son cruciales, especialmente para KNN, ya que este algoritmo se basa en distancias, y las características numéricas con escalas diferentes consiguen influir desproporcionadamente en los resultados.

Al final, un conjunto de datos bien preprocesado no solo mejora la precisión de los modelos, sino que también facilita su interpretación y análisis. Comprender las características de los conjuntos de datos es esencial para implementar correctamente K Nearest Neighbors y Naive Bayes. La elección y preparación adecuadas de los datos consiguen marcar una gran diferencia en el rendimiento de estos algoritmos de clasificación.

La disponibilidad de conjuntos de datos bien estructurados es fundamental para la implementación y evaluación de algoritmos de clasificación como K Nearest Neighbors y Naive Bayes. El conjunto de datos Iris es uno de los ejemplos más célebres en el ámbito de la estadística y el aprendizaje automático. Compilado por el botánico Edgar Anderson en 1936 y popularizado por el estadístico Ronald Fisher, este conjunto de datos incluye 150 muestras de flores de iris, distribuidas en tres especies diferentes: Iris setosa, Iris versicolor e Iris virginica (Fernández et al., 2014).

Cada muestra está representada por cuatro características numéricas: la longitud y el ancho del sépalo, así como la longitud y el ancho del pétalo. Este

conjunto de datos es ideal para la clasificación, dado que las características son fácilmente separables en un espacio bidimensional, lo que lo convierte en una excelente opción para ilustrar tanto KNN como Naive Bayes. Incluso, su tamaño relativamente pequeño facilita la visualización y el entendimiento de los resultados de clasificación.

El conjunto de datos de dígitos escritos a mano, conocido como MNIST (Modified National Institute of Standards and Technology), es un recurso ampliamente utilizado para la clasificación de imágenes. Contiene 70,000 imágenes en escala de grises de dígitos manuscritos (0-9), cada una con una resolución de 28x28 píxeles. Este conjunto de datos es particularmente valioso para evaluar algoritmos de clasificación debido a su diversidad y los desafíos que presenta en términos de variabilidad en la escritura. KNN y Naive Bayes consiguen ser aplicados para clasificar estas imágenes basándose en las características pixeladas, y su rendimiento se consigue medir a través de la precisión en la identificación correcta de los dígitos. Al respecto, la riqueza de este conjunto de datos lo convierte en un estándar de referencia para nuevas técnicas de aprendizaje automático.

El conjunto de datos de Titanic es otro clásico en el ámbito de la ciencia de datos y el aprendizaje automático. Este conjunto se basa en la infame tragedia del hundimiento del RMS Titanic en 1912 y contiene información sobre los pasajeros, incluyendo características como la edad, el sexo, la clase de billete y si sobrevivieron o no. Con 887 entradas, este conjunto de datos es particularmente atractivo para la clasificación binaria, donde el objetivo es predecir si un pasajero sobrevivió (1) o no (0). La riqueza de atributos categóricos en este conjunto de datos posibilita que tanto KNN como Naive Bayes se utilicen de manera efectiva, ya que ambos métodos consiguen manejar datos tanto numéricos como categóricos. Ahora bien, el conjunto de datos de Titanic es muy utilizado en competiciones y tutoriales, lo que lo convierte en un excelente recurso educativo.

Como se ha dicho, estos conjuntos de datos no solo son ampliamente reconocidos y utilizados en la literatura de aprendizaje automático, sino que al igual dedican una variedad de características que los hacen ideales para la implementación de algoritmos de clasificación como KNN y Naive Bayes. Su estudio y aplicación posibilitan a los investigadores y practicantes profundizar

en las capacidades y limitaciones de estos métodos, contribuyendo así al avance en el campo del análisis de datos.

La comparación entre K Nearest Neighbors (KNN) y Naive Bayes es fundamental para seleccionar el método más adecuado en función de las características específicas del problema de clasificación que se esté abordando (Arora et al., 2023). Ambos algoritmos presentan ventajas y desventajas que consiguen influir en su desempeño según el contexto de uso. No requiere un proceso de entrenamiento explícito, ya que las decisiones se toman basándose en la proximidad de los puntos de datos en el espacio de características. Esto lo convierte en un método intuitivo y accesible para aquellos que se inician en el aprendizaje automático.

No obstante, KNN siempre tiene desventajas significativas. Su rendimiento consigue verse afectado negativamente por la dimensionalidad alta de los datos, lo que se conoce como "la maldición de la dimensionalidad". En suma, KNN consigue ser computacionalmente costoso, ya que requiere calcular la distancia entre el punto de consulta y todos los puntos en el conjunto de datos, especialmente en conjuntos de datos grandes. Esto consigue resultar en un tiempo de respuesta lento en aplicaciones en tiempo real.

Por otro lado, Naive Bayes es conocido por su eficiencia y rapidez en el entrenamiento y la clasificación. Este algoritmo es especialmente eficaz en problemas donde las características son independientes unas de otras, lo que simplifica enormemente los cálculos necesarios y posibilita que el modelo se adapte rápidamente a nuevos datos. Ahora bien, la suposición de independencia condicional entre las características, que es la base del planteamiento Naive Bayes, consigue ser una limitación en situaciones donde las características están correlacionadas. Esto consigue llevar a un rendimiento subóptimo en ciertas aplicaciones. En suma, aunque Naive Bayes es efectivo con conjuntos de datos grandes, su desempeño consigue verse comprometido si los datos son escasos o si las clases están desbalanceadas.

La elección entre KNN y Naive Bayes depende en gran medida del tipo de datos y del problema de clasificación a resolver. KNN es frecuentemente utilizado en tareas donde la interpretabilidad y la precisión son cruciales, tales como en sistemas de recomendación y clasificación de imágenes. Este método es

particularmente útil cuando se dispone de un conjunto de datos relativamente pequeño y se quiere evitar hacer suposiciones sobre la distribución de los datos.

Por su parte, Naive Bayes se recomienda en situaciones donde se dispone de un conjunto de datos grande y se busca una solución rápida y eficiente, como en la clasificación de correos electrónicos (spam/no spam), análisis de sentimientos y clasificación de texto. Su capacidad para manejar datos de alta dimensionalidad con rapidez lo convierte en una elección popular en el procesamiento de lenguaje natural. Al final, tanto KNN como Naive Bayes tienen su lugar en el arsenal de técnicas de clasificación, y la elección entre uno u otro debe basarse en un análisis cuidadoso de las características del conjunto de datos y los objetivos específicos del proyecto.

En suma, la clasificación es una tarea fundamental en el campo del aprendizaje automático, y tanto K Nearest Neighbors (KNN) como Naive Bayes son métodos ampliamente utilizados debido a su simplicidad y eficacia. Los ejemplos de conjuntos de datos populares, como el conjunto de datos Iris, el de dígitos escritos a mano y el de Titanic, ilustran la amplia gama de aplicaciones que estos métodos consiguen abordar. Cada uno de estos conjuntos ofrece oportunidades únicas para aplicar KNN y Naive Bayes, destacando la versatilidad de ambos planteamientos en diferentes contextos.

Al comparar KNN y Naive Bayes, es evidente que cada uno tiene sus propias ventajas y desventajas; KNN es intuitivo y fácil de implementar, pero consigue ser costoso en términos de computación y memoria, especialmente con grandes conjuntos de datos. Por otro lado, Naive Bayes se destaca por su rapidez y eficacia en problemas de clasificación con características independientes, aunque su suposición de independencia consigue no ser válida en todos los casos. Para Madariaga et al. (2022), la elección entre KNN y Naive Bayes dependerá de las características específicas del problema en cuestión, así como de los recursos disponibles y los objetivos del análisis. Ambos métodos son herramientas valiosas en el arsenal de un científico de datos, y una comprensión profunda de los conjuntos de datos y de las condiciones en las que se aplican consigue llevar a decisiones más informadas y a un rendimiento superior en tareas de clasificación.

Conclusión

La inferencia estadística se basa en el principio de que, al estudiar una muestra, se puede obtener información valiosa sobre la población de la cual se extrajo dicha muestra. Este proceso implica el uso de modelos probabilísticos y teoremas que permiten estimar parámetros poblacionales, realizar predicciones y tomar decisiones fundamentadas. La inferencia estadística es fundamental en la investigación científica, ya que permite validar hipótesis y evaluar la efectividad de tratamientos o intervenciones en diversas disciplinas, desde la medicina hasta la economía.

En tanto, la estadística descriptiva se encarga de resumir y presentar los datos de una manera comprensible y significativa, su objetivo es describir las características de un conjunto de datos, proporcionando un panorama claro a través de gráficos, tablas y medidas numéricas. En este sentido, se limita a los datos observados y no busca hacer proyecciones o generalizaciones sobre una población más amplia.

En contraste, la estadística inferencial va más allá del mero resumen de datos, su propósito es hacer inferencias o generalizaciones sobre una población a partir de una muestra. Esto implica el uso de modelos estadísticos y teorías probabilísticas para estimar parámetros poblacionales, probar hipótesis y realizar predicciones. En resumen, mientras que la estadística descriptiva proporciona información sobre los datos en sí, la estadística inferencial permite hacer afirmaciones sobre un conjunto más amplio basado en esos datos.

¿Cuándo usar cada método?... La elección entre estadísticas descriptivas e inferenciales depende del objetivo del análisis, pues, si el objetivo es resumir y presentar datos ya recolectados, la estadística descriptiva es la opción más adecuada. Por ejemplo, un investigador que desea presentar las puntuaciones de un examen de un grupo de estudiantes puede utilizar medidas como la media y la desviación estándar para resumir el rendimiento de ese grupo específico.

Por otro lado, si se busca hacer afirmaciones sobre una población a partir de una muestra, se deberá recurrir a la estadística inferencial. En este caso, un investigador que quiera estimar la media de las puntuaciones de todos los

estudiantes de una universidad podría seleccionar una muestra representativa y aplicar técnicas inferenciales para generalizar sus hallazgos a la población total.

Al integrar ambos enfoques, los analistas de datos e investigadores pueden obtener una comprensión más profunda de estos, lo que permite tomar decisiones más informadas y basadas en evidencia. A medida que la cantidad de datos disponibles continúa creciendo en nuestra sociedad actual, la habilidad de aplicar correctamente estos métodos estadísticos se vuelve cada vez más esencial, no solo en el ámbito académico, sino también en sectores como la salud pública, la economía, la educación, la medicina experimental y muchos otros. En definitiva, dominar la estadística es una competencia clave que empodera a los individuos y organizaciones para enfrentar la fiabilidad en la creciente demanda de análisis de datos masivos en investigaciones de tipo experimental o no.

En conclusión, la medición de la fiabilidad es un proceso esencial que debe ser cuidadosamente considerado por los investigadores, la elección entre el alfa de Cronbach y los índices de Guttman, así como el entendimiento de sus implicaciones, consigue influir significativamente en la interpretación de los resultados y en la validez general de la investigación. Por lo tanto, es fundamental que los profesionales del campo se mantengan informados sobre las mejores prácticas y tratamientos actuales en la evaluación de la fiabilidad, asegurando así la calidad y la robustez de sus estudios.

Bibliografía

Arora, K., Pathak, S., & Dieu Linh, N.T. (2023). Comparative Analysis of K-Nn, Naïve Bayes, and logistic regression for credit card fraud detection. *Ingeniería Solidaria*, *19*(3), 1-22. https://doi.org/10.16925/2357-6014.2023.03.05

Ashour, W., & Sunoallah, S. (2011). Multi Density DBSCAN. In: Yin, H., Wang, W., Rayward-Smith, V. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2011. *Lecture Notes in Computer Science,* 6936. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23878-9_53

Batanero, C. (2001). *Didáctica de la estadística.* Granada: Universidad de Granada

Batanero, C., y Díaz, C. (2011). *Estadística con proyectos.* Granada: Universidad de Granada

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp 785–794). San Francisco: ACM Digital Library

Contento Rubio, M.R. (2019). *Estadística con aplicaciones en R.* Bogotá: Universidad de Bogotá Jorge Tadeo Lozano

Dagnino, J. (2014). Intervalos de confianza. *Rev. chil. anest.,* 43(2), 129-133. https://doi.org/10.25237/revchilanestv43n02.11

Drucker, H., Burges, C.J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support Vector Regression Machines. *Advances in Neural Information Processing Systems.* 9, 155-161

Engelhard, G. (2005). Guttman Scaling. En *Encyclopedia of Social Measurement* (pp. 167-174). London: Elsevier

Fau, C., y Nabzo, S. (2020). Metaanálisis: bases conceptuales, análisis e interpretación estadística. *Revista mexicana de oftalmología*, *94*(6), 260-273. https://doi.org/10.24875/rmo.m20000134

Fernández Ropero, R.M., Aguilera Aguilera, P., Fernández, A., y Rumí, R. (2014). Redes bayesianas: una herramienta probabilística en los modelos de distribución de especies. *Ecosistemas,* 23(1), 54-60

Font, X. (2019). *Técnicas de clustering.* Barcelona: Editorial UOC

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). Dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software*, *91*(1), 1–30. https://doi.org/10.18637/jss.v091.i01

Kramer, O. (2013). K-Nearest Neighbors. In: Dimensionality Reduction with Unsupervised Nearest Neighbors. *Intelligent Systems Reference Library*, 51. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38652-7_2

Kuter, K. (2025). *Math 345-Probability (Kuter).* California: LibreTexts

Lamfre, L., Perren, J., & Bramardi, S. (2023). Análisis de Correspondencias Múltiple Condicionada. Una aplicación al estudio de la calidad de vida según clase social en la Argentina de inicios del milenio. *SaberEs*, *15*(2), 157–175. https://doi.org/10.35305/s.v15i2.272

Luzuriaga, H.A., Espinosa, C.A., Haro, A.F., y Ortiz, H.D. (2023). Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS. *LATAM Revista Latinoamericana De Ciencias Sociales Y Humanidades*, *4*(4), 596– 607. https://doi.org/10.56712/latam.v4i4.1242

Madariaga Fernández, C.J., Lao León, Y.O., Curra Sosa, D.A., y Lorenzo Martín, R. (2022). Empleo de algoritmos KNN en metodología multicriterio para la clasificación de clientes, como sustento de la planeación agregada. *Retos de la Dirección*, *16*(1), 178-198

Martínez Román, J.A. (2009). *Análisis y modelización del comportamiento innovador de las empresas. Una aplicación a la provincia de Sevilla.* Sevilla: Consejo Económico y Social de Andalucía

Martinson, D.G. (2018). Teoría de la probabilidad. En *Métodos cuantitativos de análisis de datos para las ciencias físicas y la ingeniería* (pp. 15-61). Cambridge: Cambridge University Press

Meneses, J., Barrios, M., Bonillo, A., Cosculluelka, A., Lozano, L.M., Turbani, J., y Valero, S. (2013). *Psicometría.* Barcelona: Editorial UOC

Mias, C.D., y Tornimbeni, S. (2020). *Metodología, estadística aplicada e instrumentos en Neuropsicología: Guía práctica para investigación.* Córdoba: Editorial Brujas

Miyamoto, S. (2022). *Theory of Agglomerative Hierarchical Clustering.* Singapore: Springer Singapore. https://doi.org/10.1007/978-981-19-0420-2

Mnyanghwalo, D., Kundaeli, H., Kalinga, E. y Hamisi, N. (2020). Enfoques de aprendizaje profundo para la detección y clasificación de fallas en la red de distribución secundaria eléctrica: Comparación de métodos y de la precisión de redes neuronales recurrentes. *Cogent Engineering* , *7* (1). https://doi.org/10.1080/23311916.2020.1857500

Moors, J.J.A. (1988). A Quantile Alternative for Kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *37*(1), 25–32. https://doi.org/10.2307/2348376

Oviedo, H.C., & Campo-Arias, A. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista Colombiana de Psiquiatría*, *34*(4), 572-580

Posada, G.J. (2016). *Elementos básicos de estadística descriptiva para el análisis de datos.* Medellín: Fundación Universitaria Luis Amigó

Prieto, G., & Delgado, A.R. (2010). Fiabilidad y validez. *Papeles del Psicólogo,* 31(1), 67-74. Disponible en: https://www.redalyc.org/pdf/778/77812441007.pdf

Quero, M., y Inciarte, K.(2012). Clasificación de las Técnicas Estadísticas Multivariantes. *Telos,* 14(2), 275-286

Quevedo, F. (2011). Medidas de tendencia central y dispersión. *Medwave,* 11(3), 1-6. http://doi.org/10.5867/medwave.2011.03.4934

Reynolds, D. (2009). Modelos de mezcla gaussiana. En: Li, S. Z., Jain, A. (eds.) Enciclopedia de Biometría. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_196

Roco-Videla, Á., Landabur-Ayala, R., Maureira-Carsalade, N., y Olguin-Barraza, M. (2023). ¿Cómo determinar efectivamente si una serie de datos sigue una distribución normal cuando el tamaño muestral es pequeño?. *Nutrición Hospitalaria*, *40*(1), 234-235. https://dx.doi.org/10.20960/nh.04519

Rodríguez, C.R., Breña. J.L. y Esenarro, D. (2021). *Las variables en la metodología de investigación científica.* Alicante: Editorial Área de Innovación y Desarrollo. https://doi.org/10.17993/IngyTec.2021.78

Rodríguez-Pérez, R., & Bajorath, J. (2022). Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *Journal of computer-aided molecular design*, *36*(5), 355–362. https://doi.org/10.1007/s10822-022-00442-9

Sánchez-Solis, Y., Raqui-Ramirez, C. E., Huaroc-Ponce, E. J., y Huaroc-Ponce, N. M. (2024). Importancia de Conocer la Normalidad de los Datos Utilizados en los Trabajos de Investigación por Tesistas. *Revista Docentes 2.0*, *17*(2), 404–413. https://doi.org/10.37843/rted.v17i2.554

Sinha, P., Calfee, C.S., & Delucchi, K.L. (2021). Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. *Critical care medicine*, *49*(1), e63–e79. https://doi.org/10.1097/CCM.0000000000004710

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, S., Bandeen-Roche, K., Béland, F., & Bergman, H. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of clinical epidemiology*, *63*(6), 638–646. https://doi.org/10.1016/j.jclinepi.2009.08.008

Taha, K. (2023). Semi-supervised and un-supervised clustering: A review and experimental evaluation. En *Information Systems* (e102184). London: Cambridge Elsevier

Valova, I., Gueorguieva, N., Mai, T., y Chen, R. (2024). Agrupamiento difuso en espacios de alta dimensión: Visualización y métricas de rendimiento. *Revista Internacional de Sistemas de Ingeniería Inteligente y Basados en el Conocimiento*, 28(2), 313-333. https://doi.org/10.3233/KES-221614

van Maarseveen, H. (2023*). XGBoost: The Ultimate Guide to Extreme Gradient Boosting.* Tokyo: Henri van Maarseveen

Wade, C., y Glynn, K. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python.* Birmingham: Packt Publishing

Williams, C.K. (2002). On a connection between kernel PCA and metric multidimensional scaling. *Machine Learning,* 46(1), 11-19

Yadav, P., y Dhull, A. (2024). *Una Técnica de Clustering Jerárquico Eficiente para el Diagnóstico Médico.* México: Editorial Nuestro Conocimiento

Esta edición de *"Métodos de análisis estadístico: Desde lo descriptivo hasta lo inferencial"* se culminó en la ciudad de Colonia del Sacramento en la República Oriental del Uruguay el 02 de mayo de 2025



MÉTODOS DE ANÁLISIS ESTADÍSTICO: DESDE LO DESCRIPTIVO HASTA LO INFERENCIAL

ISBN: 978-9915-698-07-6
9 789915 698076